“Tables are tricky”. Testing Text Encoding Initiative (TEI) Guidelines for FAIR upcycling of digitised historical statistics.[1]

Gabi Wüthrich

In 2121, nothing is as it once was: a nasty virus is keeping the world on tenterhooks – and people trapped in their own four walls. Older people remember their grandmothers' horror stories of missing toilet paper and banana-flavoured coal briquettes. Pandemics had already turned the lives of many people upside down 100 and 200 years earlier. In the depths of the metaverse, contemporaries are searching for data to compare the frightening death toll of the current killer virus with its predecessors during the Covid-19 pandemic and the «Spanish flu». There is an incredible amount of statistical material on the Covid-19 pandemic in particular, but annoyingly, this is only available in obscure data formats such as .xslx in the internet archives. They can still be opened with the usual text editors, but their structure is terribly confusing and unreadable with the latest statistical tools. If only those digital hillbillies in the 2020s had used a structured format that not only long-outdated machines but also people in the year 2121 could read...

Admittedly, very few epidemiologists, statisticians and federal officials are likely to have considered such future scenarios during the pandemic years. Quantitative social sciences and the humanities, including medical and economic history, but also memory institutions such as archives and libraries, should consciously consider how they can sustainably preserve the flood of digital data for future generations. Thus, the sustainable processing and storage of statistical printed data from the time of the First World War makes it possible to gain new insights into the so-called «Spanish flu» e. g. in the city of Zurich even today. The publications by the Statistical Office of the City of Zurich, which were previously only available in “analog” paper format, have been digitized by the Zentralbibliothek (Central Library, ZB) Zurich as part of Joël Floris' Willy Bretscher Fellowship 2022/2023.[2] This project paper has been written in the context of this digitisation project, as issues regarding digital recording, processing, and storage of historical statistics have always occupied quantitative economic historians “for professional reasons”.

The basic idea of this paper is to prepare tables with historical health statistics in a sustainable way so that they can be easily analysed using digital means. The aim was to capture the statistical publications retro-digitized by the ZB semi-automatically with OCR in Excel tables and to prepare them as XML documents according to the guidelines of the Text Encoding Initiative (TEI), a standardized vocabulary for text structures. To do this, it was first necessary to familiarise with TEI and its appropriate modules, and to apply them to a sample table in Excel. To be able to validate the Excel table manually transferred to XML, I then developed a schema based on the vocabularies of XML and TEI. This could then serve as the basis for an automated conversion of the Excel tables into TEI-compliant XML documents. Such clearly structured XML documents should ultimately be relatively easy to convert into formats that can be read into a wide variety of visualisation and statistical tools.

[1] This paper is a translation (with deepL) and revision of a project paper titled “From book to machine. XML as a platform-independent, machine-readable table format” written by the author as a contribution to modules 2 and 3 of the Certificate of Advanced Studies (CAS) UZH in Data Management and Information Technologies. It was supervised and accepted by Dr. Mathias Müller, University of Zurich, in Spring 2022.

[2] Floris, Joël. Die Spanische Grippe in Zürich erzählen. Willy-Bretscher-Fellow 2022/23. Juni 2023. <https://www.zb.uzh.ch/de/zuerich/die-spanische-grippe-zuerich-erzaehlen> [assessed on 22.11.2023].

Data description

A table from the monthly reports of the Zurich Statistical Office serves as an example data set. The monthly reports were digitised as high-resolution pdfs with underlying Optical Character Recognition (OCR) by the Central Library's Digitisation Centre (DigiZ) as part of the Willi Bretscher Fellowship project, and were initially made available to me by Joël Floris. They are now available on the ZB's Zurich Open Platform (ZOP), including detailed metadata information.[3] They were published by the Statistical Office of the City of Zurich as a journal volume under this title between 1908 and 1919, and then as «Quarterly Reports» until 1923.[4] The monthly reports each consist of a 27-page table section with individual footnotes, and conclude with a two-page explanatory section in continuous text.

For this study, the data selection is limited to a table for the year 1914 and the month of January. In connection with Joël Floris' project, which aims at obtaining quantitative information on Zurich's demographic development during the «Spanish flu» from the retro-digitisation project, it was obvious to focus on tables with causes of death. The corresponding table number 12 entitled «Die Gestorbenen (in der Wohnbev.) nach Todesursachen und Alter» («The Deceased (in the Resident Pop.) by Cause of Death and Age») can be found on page seven of the monthly report. It contains monthly data on causes of death, broken down by age group and gender, as well as comparative figures for the same month of the previous year. The content of this table is to be prepared below in the form of a standardized XML document with an associated schema that complies with the TEI guidelines.

Methods for capturing historical tables in XML

The source of inspiration for this project paper was a pioneering research project originally based at the University of Basle. In the research project, the annual accounts of the city of Basle from 1535 to 1610 were digitally edited. Technical implementation was carried out by the Center for Information Modeling at the University of Graz.[5] Based on a digital text edition prepared in accordance with the TEI standard, the project manages to combine facsimile, web editing in HTML, and table editing via an RDF (Resource Description Framework[6]) and XSLT (eXtensible Stylesheet Language Transformations[7]) in an exemplary manner.[8] The edition thus allows users to compile their own selection of booking data in a «data basket» for subsequent machine-readable analysis. In an accompanying article, project team member Georg Vogeler describes the first-time implementation of a numerical evaluation and how «even extensive holdings can be efficiently edited digitally».[9] However, as mentioned, the central basis for this is XML processing of the corresponding tabular information based on the TEI standard.

---

[3] See Statistisches Amt der Stadt Zürich. Monats-Berichte des Statistischen Amtes der Stadt Zürich 1914. https://doi.org/10.20384/zop-869.

[4] See catalogue entry in Swisscovery: https://uzb.swisscovery.slsp.ch/permalink/41SLSP_UZB/1j14kmp/alma990009955150205508. The volumes prior to 1908 appear to have been published in the Tagblatt der Stadt Zürich. The ZB has transferred their entire holdings to the Speicherbibliothek, but has since digitized them as part of the Bretscher fellowship, and made them accessible on ZOP in the statistics section (https://www.zop.zb.uzh.ch/handle/123456789/829).

[5] See Burghartz. Jahrrechnungen der Stadt Basel. http://gams.uni-graz.at/context:srbas.

[6] See https://www.w3schools.com/xml/xml_rdf.asp.

[7] See https://www.w3schools.com/xml/xsl_intro.asp.

[8] See Sagelsdorff, Vogler und Burghartz. Jahrrechnungen der Stadt Basel. Projektbeschreibung. http://gams.uni-graz.at/context:srbas?mode=projekt, here: Digitale Edition/Datenmodell.

[9] Vogeler, Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert? https://zfdg.de/sb001_007#vogeler_edition.

XML as a method standard

XML stands for «eXtensible Markup Language», can be read by both humans and machines and serves as the basis for storing and processing data. An XML document mainly consists of elements, attributes, and text, usually begins with an XML declaration and can contain comments. Elements have a start and an end tag, and can contain attributes, child elements and text, but can also be empty. Their names must fulfil certain structural requirements. Attributes, on the other hand, have a name and a value enclosed in quotation marks. XML documents should be well-formed, i. e. contain an introductory XML declaration and a single root element. In addition, their elements must be precisely nested hierarchically, and their names and attributes must be implemented correctly. For the contents of these general XML documents to have meaning, they must be linked to a vocabulary that defines specific rules for the elements. To do this, the names of the elements and attributes can be linked to specific namespaces that are stored in an XML schema and defined there. Elements can be described in these schemas either as predefined simple types, or as complex types to be defined further. Using the schemas with the help of text editors, we can determine whether an XML document is valid in relation to the defined namespace schema, i.e. a defined vocabulary.

TEI as a methodological basis for digital text forms

The Text Encoding Initiative (TEI) provides a comprehensive vocabulary in the XML standard[10] with which texts can be structured in a standardised way. Lou Burnard, one of the TEI pioneers, describes the purpose of the TEI as follows: «Its purpose is to provide guidelines for the creation and management in digital form of every type of data created and used by researchers in the Humanities, such as source texts, manuscripts, archival documents, ancient inscriptions, and many others. »[11] Accordingly, the current TEI is characterized by what texts looked like in the past but does not want to exclude future forms of text. Nevertheless, it should treat all digital documents equally, regardless of whether they are «digitally born» or not.[12] According to Burnard, the preparation of texts in TEI has three decisive advantages: it focuses on the meaning and not on the layout format of texts, is software-independent, and is supported by an open scientific community.[13]

In April 2022, an update of the latest version (4.4.0) of the TEI guidelines was published, which I used as the basis for the project. The guidelines contain 23 chapters with information on the TEI elements and modules as well as on how to apply them[14], including a chapter on the preparation of tables, formulas, graphics, and music.[15] As a pdf document, the guidelines comprise around 2000 pages, but only about four of these deal with the presentation of tables. Even the introduction to Chapter 14 is being rather cautious with regard to TEI application for table formats, warning that layout and presentation details are more important in table formats than in running text, and that they are already covered more comprehensively by other standards and should be prepared accordingly in these notations.[16]

On asking the TEI-L mailing list whether it made sense to prepare historical tables with the TEI table module, the answers were rather reserved.[17] Only the Graz team remained optimistic that TEI could be used to process historical tables, albeit in combination with an RDF including a corresponding

[10] See Appendix A. XML as a method standard.
[11] Burnard, What is the Text Encoding Initiative?, Introduction. https://books.openedition.org/oep/679.
[12] Burnard, What is the Text Encoding Initiative?, The TEI and XML. https://books.openedition.org/oep/680.
[13] Burnard, What is the Text Encoding Initiative?, Introduction. https://books.openedition.org/oep/679.
[14] P5: Guidelines for Electronic Text Encoding and Interchange, https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.
[15] Kap. 14, https://tei-c.org/release/doc/tei-p5-doc/en/html/FT.html.
[16] Kap. 14, https://tei-c.org/release/doc/tei-p5-doc/en/html/FT.html.
[17] https://listserv.brown.edu/cgi-bin/wa?A1=ind2206&L=TEI-L#20.

ontology.[18] Christopher Pollin also provided github links via TEI to the DEPCHA project currently underway, for which they are developing an ontology for annotating transactions in historical account books.[19]

Table structure in TEI-XML

Basically, the TEI schema treats a table as a special text element consisting of line elements, which in turn contain cell elements.[20] This basic structure was used to code Table 12 from 1914, which I transcribed manually as an Excel file (see Appendix B. Structural features Table 12). Because exact formatting including precise reproduction of the frame lines is very time-consuming, the frame lines in the project work only served as structural information and are not included as topographical line elements as TEI demands.[21] Long dashes, which correspond to zero values in the source, are interpreted as empty values in the TEI-XML. I used the resulting worksheet as the basis for the TEI-XML annotation, in which I also added some metadata.

In order to code the XML structure in my TEI conforming XML version of the Excel file correctly, I used Visual Studio Code with the XML Language Support from Red Hat in version v0.20.0, and the necessary Java extension.[22] The introductory, non-mandatory XML declaration is kept as general as possible as is usual in XML notation. In contrast, the start tag of a TEI-compliant document includes namespace definitions from TEI and W3. For the project work, I created a correspondingly adapted local schema in a second step, which I refer to accordingly in the XML document.

A document encoded according to TEI-XML must first contain a TEI header with some mandatory elements that provide metadata for the text body that is subsequently digitally processed. The TEI header basically comprises five «parent» elements.[23] However, only the first of these, the file description with the fileDesc tag, is mandatory. Accordingly, I have limited myself to the child elements of fileDesc in the TEI header in the project work. This in turn contains three mandatory elements, namely information on the title of the digital work (titleStmt), the type of digital publication (publicationStmt) and the original source (sourceDesc) on which the digital work is based and which contains detailed bibliographic metadata. In addition, there are other elements that are not absolutely necessary, such as the edition statement (editionStmt), which indicates the edition on which this digital publication is based – in this specific case, the table digitized by the ZB.

After the TEI header, the actual text body begins with the «text» element, which is immediately followed by the «body» with the start section «div» due to the restriction to a single table.[24] Suitable heading («head») elements are the title of the table, the table number as a note and the «date» of the table (see Appendix B. Structural features Table 12). The first table row contains the column headings and is assigned the role attribute «label» accordingly. The third-last cell of each row contains the row total, which I have given the attribute «ana» for analysis and the value «#sum» for total, following the example of the Basle Edition. The first cell of each row again names the cause of death and must therefore also be labelled with the role attribute «label». The second last row shows the sum of the

[18] Christopher Pollin's reply via the TEI-L mailing list (see https://listserv.brown.edu/cgi-bin/wa?A1=ind2206&L=TEI-L#20) resulted in a personal email contact, which I will be happy to forward on request.
[19] See https://github.com/MEDEAEditions/DEPCHA.
[20] See Kapitel 14 der TEI-Guidelines. https://tei-c.org/release/doc/tei-p5-doc/en/html/FT.html.
[21] See Kapitel 11 der TEI-Guidelines. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHZLAB.
[22] https://marketplace.visualstudio.com/items?itemName=redhat.vscode-xml. Entgegen der eigenen Beschreibung des Tools funktioniert es in Windows weiter nur mit der Java-Erweiterung.
[23] Cited from https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html. See Appendix A. TEI header elements.
[24] See https://tei-c.org/release/doc/tei-p5-doc/en/html/DS.html.

current monthly table, which is why it is given the «#sum» attribute for all respective cells. Finally, the last line shows the total for the previous year's month. It is therefore not only marked with the sum attribute, but also with a date in the label cell. A potential confounding factor for later calculations is the row «including diarrhea», which further specifies diseases of the digestive organs but must not be included in the column total. Accordingly, it is provided with another analytical attribute called «#exsum». As each cell in the code represents a separate element, the »digitally upcycled table 12 in XML format ultimately extends over a good 550 lines of code (see textbox below).

```
                <head>
                    <title>Die Gestorbenen (in der Wohnbevölkerung) nach Todesursachen und
Alter</title>
                    <note n="Tabelle 12"></note>
                    <date>"Januar 1914"</date>
                </head>1914_12_AdjSchema.xml
                    <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 -->
```

TEI schema for table

The next step was to validate my TEI formatted XML version of the table against a TEI-based schema. Because the tei_all schema provided by TEI comprises a good 23,000 lines of code and is correspondingly overwhelming, it had be condensed to the few TEI elements I used in the XML document. To do this, TEI provides a tool called Roma[25], which can be employed to create schemas according to the prevalent W3 standard. I chose the first variant of the «build up», in which I selected the two modules «figures» (14) and «analysis» (17) in addition to the smallest recommended

[25] Roma: generating customizations for the TEI. https://roma2.tei-c.org/.

schema[26]. The superordinate schema language is English, while a German translation is available for for the descriptions used in the schema, i. e. for the output language.[27]

However, the main schema document (in the format of an .xsd-File) generated by Roma still contains more than 8000 lines of code. To bring this into a somewhat more compact schema form for the project work, I first copied its header, before systematically searching for elements used in the TEI-XML version of table 12 and copied them all into a new XML schema. The advantage of copying the elements is not only to design them according to the TEI specifications, but also to retain the descriptions of the respective elements structured with «annotation», as they contribute a great deal to understanding the structure of TEI.

In this minimal version of the schema, it was sufficient to convert various elements that are actually defined as complex in the original TEI schema into simple types that just have to contain a text element (e.g. <title> with the restriction «string»).[28] The references used in the original TEI code to definitions of child elements with the tag <ref> make the code leaner, so I have adopted them accordingly. Similar to the TEI-XML file, the elements of the TEI header need to be defined in the schema. The schema then allows to set indicators for the minimum occurrences (minOccurs) for the elements defined, i. e. to make them mandatory in the XML files according to TEI standards. Thus, certain header and body elements need to be present to have a clean XML complying to TEI. In addition, several adjustments to elements are included in the schema to do justice to more complex contents of a cell in the minimal schema. In total, the .xsd-schema I created to check conformity of the XML version of table 12 with TEI rules comprises of around 530 lines of code.

Challenges and problems

An initial problem already arose during the OCR-based digitisation. The Central Library's OCR software, which specializes in continuous text, simply failed to capture the text in the tables (see Appendix C. Original Output OCR). I therefore first had to transcribe the table by hand, which can be prone to errors. In principle, however, it is irrelevant in TEI in which format the original text was created.[29] The potential for errors when transferring Excel data into the «original» XML is also high, especially if the table is complex and/or detailed. Ideally, i. e. with a clean OCR table, it ought to be possible to export OCR content in pdfs to XML.

Due to the extremely short instructions for table preparation in TEI, I underestimated the variety of different text components that TEI offers. The complexity of TEI is not clear from the rough overview of the individual chapters and their introductory descriptions. This only became clear while adjusting table 12 to TEI standards. By becoming accustomed to TEI, its limitations regarding table preparation also became more evident: It is fundamentally geared towards structuring continuous text rather than text forms, where the structure or layout also indicates meaning, as is the case with tables.

The conversion of the sample table into XML and the preparation of an associated TEI schema, which is reduced to the elements present in the sample document, yet remains valid with the TEI standard, proved to be time-consuming code work. The restriction to the elements contained in the sample XML required countless adjustments to the TEI standard elements, which are defined via complex element

[26] In addition, I included the orgName element from module 13 (Names, Dates, People, and Places) https://www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-org.html), which I adopted from the tei_all schema.
[27] In the end, the Roma tool generates a ZIP file with three documents, two of which contain short supplementary attribute definitions for elements that I did not use in my minimal example.
[28] To keep track of these adjustments, the respective code lines include a comment «Changed to simple Type».
[29] See chapter Iv. About These Guidelines. https://tei-c.org/release/doc/tei-p5-doc/en/html/AB.html.

and attribute groups. On the other hand, these customisation steps helped me to better understand the basic idea and structure of TEI. Nevertheless, both the sample XML and the local schema each comprise over 500 lines of code – and this basically for only a single – though complex – table with a few metadata.

In addition, the project outline had also considered an automated conversion of Excel into a TEI-compatible XML document. However, when I tried to use the XML document created in TEI as a basis for an automated XML mapping in Excel, I failed miserably. The extremely comprehensive and complex TEI schema on which my XML document is based is not suitable for implementation in Excel. As a result, I had to prepare an XML table schema that was as general as possible (see Appendix D. Automated conversion from Excel to XML), which in a second step could automatically convert the Excel tables into XML, thus reducing error potential of the XML conversion.

Ideas for Project Expansion

Because, as mentioned, the OCR output of the tables in this case is not usable, it should now be crucial for any digitisation project to achieve high-quality OCR of the retro-digitised tables. Table recognition is definitely an issue in economic history research, and there are several open source development tools around on Github, which yet have to set a standard, however.[30]

Ideally, the tables recognized in this way would then provide better text structures in the facsimile. With the module for the transcription of original sources, TEI offers extensive possibilities for linking text passages in the transcription with the corresponding passages in the facsimiles. Such links could ideally be used as training data for text recognition programs to improve their performance in the area of table recognition. Other TEI elements that lend structure to the table, such as the dividing lines and the long dashes for the empty values, could also serve as such structural recognition features.

Additional important TEI elements such as locations and gender would further increase the content of the TEI XML format. Detailed metadata, as e.g. provided by the retro-digitized version of the ZOP, can be easily integrated into the TEI header area «xenodata». Finally, in view of the complex structure of the tables, it is essential to understand and implement XSLT (eXtensible Stylesheet Language Transformation) for automated structuring, and as a basis for RDF used e.g. by the Graz team.

Conclusion

So far, tables seem to have had a shadowy existence within the Text Encoding Initiative (TEI) – or, as TEI pioneer Lou Burnard remarked in the TEI mailing list on behalf of my question whether TEI processing of tables made sense: «Tables are tricky». The main reason for this probably lies in the continuous text orientation of existing tools and users, who are also less interested in numerical formats.

In principle, however, preparation according to the TEI standard offers the opportunity to think conceptually about the function of tabularly structured data and to make changes, e.g. in serial sources such as statistical tables, comprehensible. The clearly structured text processing of TEI could provide a basis for improving the still rather poor quality of text recognition programs when recording tables. And a platform-independent, non-proprietary data structure such as XML would be almost indispensable for the sustainable long-term archiving of «digitally born» statistics, which have

[30] See e.g. a tool developed by the University of Bern (https://github.com/ub-unibe-ch/ds-pytools/blob/main/OCR/segmentation_for_OCR/table_segmentation_%26_data_field_extraction_BZ.ipynb) or one by data scientist Rebecca Zhao (https://beczzzhao.github.io/CVHistorical/).

experienced a boom in recent years, and especially during the pandemic. After all, our descendants should also be able to access historical statistics during the next one.

References

Burnard, Lou. What is the Text Encoding Initiative? How to add intelligent markup to digital resources. Marseille: OpenEdition Press, 2014. <https://books.openedition.org/oep/426>.

Burnard, Lou und Bauman, Syd. P5: Guidelines for Electronic Text Encoding and Interchange, Version 4.4.0, 2022. <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html> [abgerufen ab 22. März 2022].

Floris, Joël. Digitizing demographic data form Zurich, 1910-1925, to quantify and contextualize the "Spanish flu". Antrag zur Ausschreibung Willy-Bretscher-Fellowships 2022/2023 der Zentralbibliothek Zürich.

Müller, Mathias. Materialien zum CAS_DMIT_22_XML. <https://drive.google.com/drive/folders/1hCSZdMM9mGG--YfnSNMkM3RjciJ-e4Sh> [abgerufen ab 18. Februar 2022].

Burghartz, Susanna (Hg.). Jahrrechnungen der Stadt Basle 1535 bis 1610 – digital. Unter Mitarbeit von Calvi, Sonja, Meili, Lukas, Sagelsdorff, Jonas, und Vogler, Georg. Basle/Graz 2015. <http://gams.uni-graz.at/context:srbas> [abgerufen ab 25. Februar 2022].

Statistisches Amt der Stadt Zürich. Monats-Berichte des Statistischen Amtes der Stadt Zürich 1914. Zürich: Buchdruckerei Berichthaus, 1915. <https://doi.org/10.20384/zop-869> [abgerufen am 13. Juni 2022].

Statistisches Amt der Stadt Zürich. Monats-Berichte des Statistischen Amtes der Stadt Zürich 1918. Zürich, 1919. <https://doi.org/10.20384/zop-871> [abgerufen am 13. Juni 2022].

TEI-L mailing list. TEI for tables. <https://listserv.brown.edu/cgi-bin/wa?A1=ind2206&L=TEI-L#20> [abgerufen am 17. Juni 2022], sowie persönliche Mails mit Christopher Pollin und Werner Stangl von der Uni Graz (Details auf Anfrage).

Vogeler, Georg. Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?. *In:* Baum, Constanze und Stäcker, Thomas (Hg.). Grenzen und Möglichkeiten der Digital Humanities. 2015 (Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). <10.17175/sb001_007> [abgerufen am 14. Juni 2022].

W3Schools. Various tutorials on XML. <https://www.w3schools.com/> [accessed March 22, 2022].

Appendix

A. TEI header elements

The TEI header comprises five «parent» elements:

1. a file description with the tag fileDesc: This contains the complete bibliographic description of the computer file and is the only one of the five elements that must be specified in the XML file.

2. an encoding description with the tag encodingDesc: This describes the relationship between the electronic text and the original source.

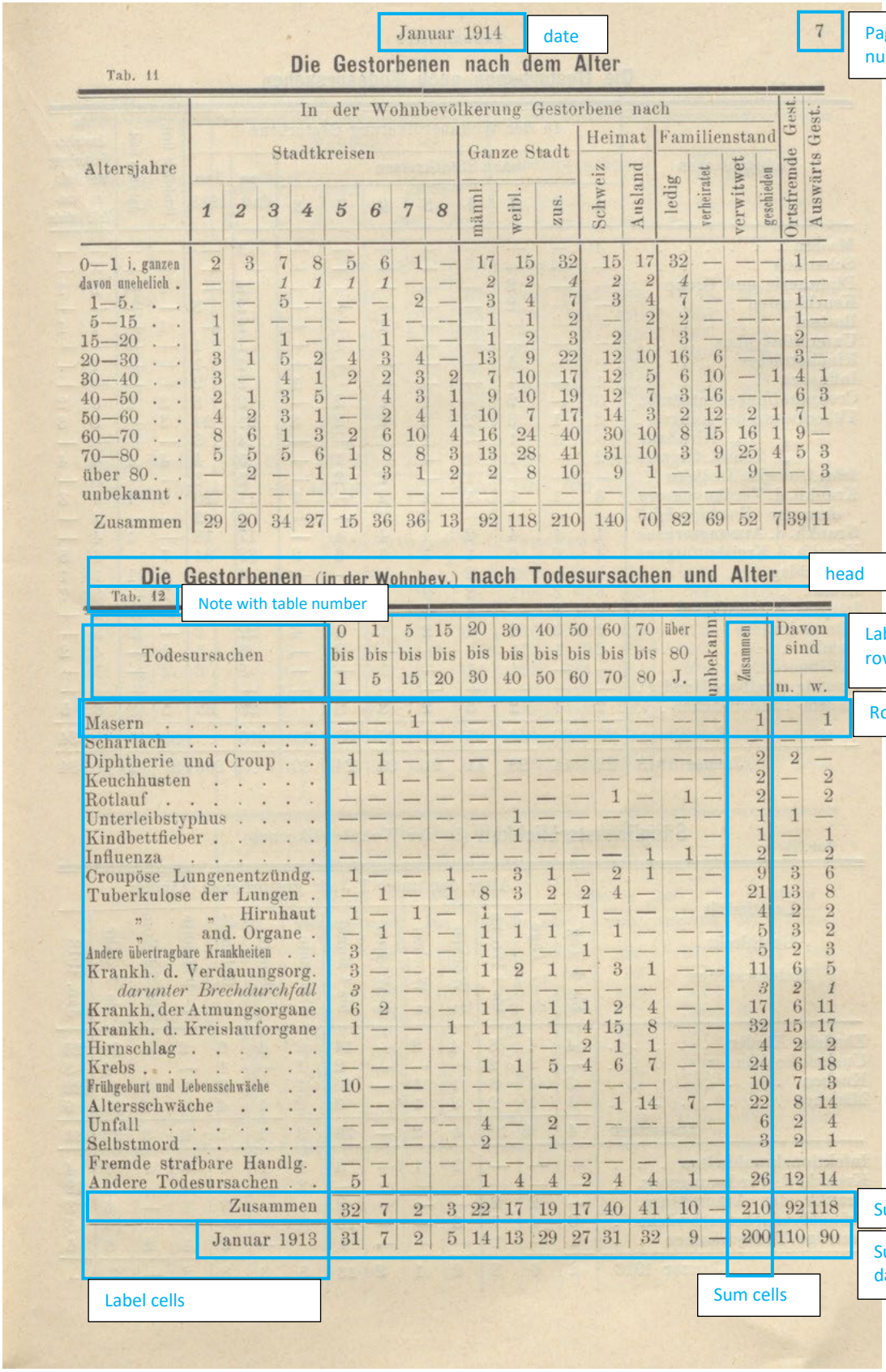
3. a text profile with the tag profileDesc: It contains classifying and contextual information about the text.

4. a container element with the tag xenoData: This allows the simple integration of metadata that has not been encoded according to TEI. This contains the references to the corresponding MARC records, which are particularly interesting from a library perspective.

5. a revision list with the revisionDesc tag: This allows you to document changes in the electronic recording process.

B. Structural features Table 12

Monthly Report January 1914



Januar 1914 — date

7 — Page number

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 — head

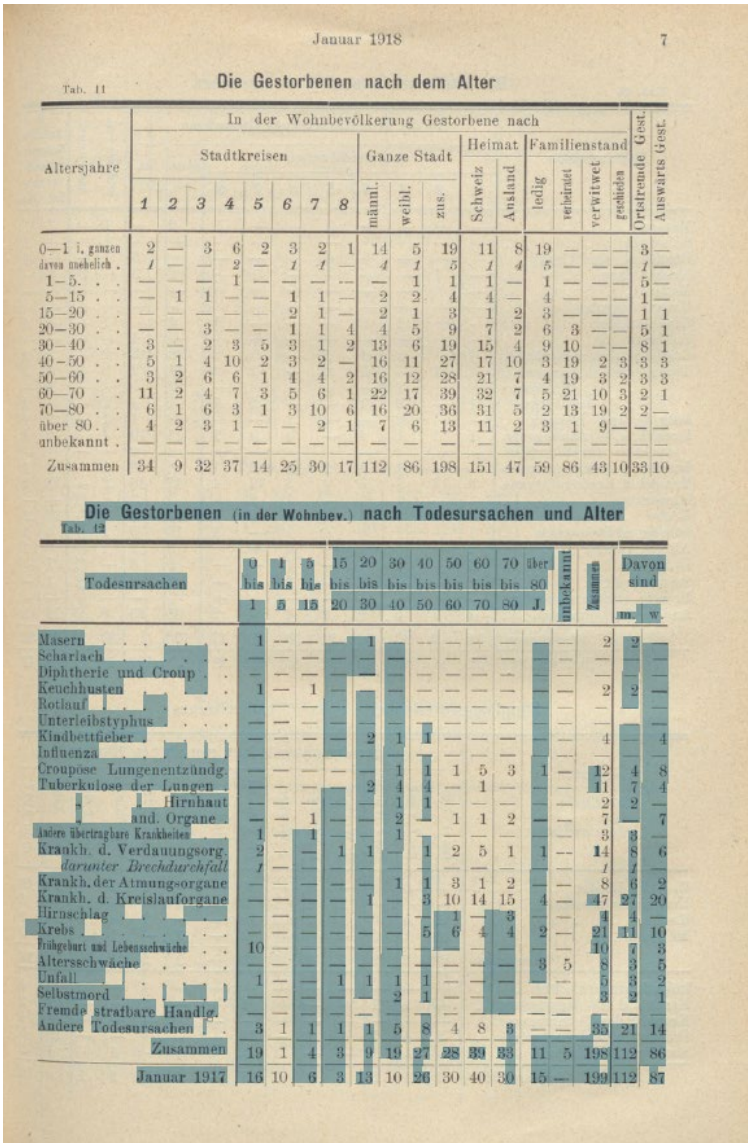
Tab. 12 — Note with table number

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	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

Label row — Row 1 — Sum row — Sum row date 1913 — Label cells — Sum cells

C. Original Output OCR

Selected in the pdf-file



Januar 1918 7

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 J. ganzen	2	—	3	6	2	3	2	1	14	5	19	11	8	19	—	—	—	3	—
davon uneheliche	1	—	—	2	—	1	1	—	4	1	5	1	4	5	—	—	—	1	—
1—5. . .	—	—	—	1	—	—	—	—	—	1	1	1	—	1	—	—	—	5	—
5—15 . .	—	1	1	—	—	1	1	—	2	2	4	4	—	4	—	—	—	1	—
15—20 . .	—	—	—	—	—	2	1	—	2	1	3	1	2	3	—	—	—	1	1
20—30 . .	—	—	3	—	—	1	1	4	4	5	9	7	2	6	3	—	—	5	1
30—40 . .	3	—	2	3	5	3	1	2	13	6	19	15	4	9	10	—	—	8	1
40—50 . .	5	1	4	10	2	3	2	—	16	11	27	17	10	3	19	2	3	3	3
50—60 . .	3	2	6	6	1	4	4	2	16	12	28	21	7	4	19	3	2	3	3
60—70 . .	11	2	4	7	3	5	6	1	22	17	39	32	7	5	21	10	3	2	1
70—80 . .	6	1	6	3	1	3	10	6	16	20	36	31	5	2	13	19	2	2	—
über 80 . .	4	2	3	1	—	—	2	1	7	6	13	11	2	3	1	9	—	—	—
unbekannt .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Zusammen	34	9	32	37	14	25	30	17	112	86	198	151	47	59	86	43	10	33	10

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	—	—	—	—	—	—	—	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

ORC 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

D. Automated conversion from Excel to XML

Using the Developer tab in Excel, tables could in principle be exported relatively easily as XML by inserting a corresponding source XML as the basis for an XML mapping. However, as Excel uses the underlying schema for the mapping in the source XML, direct import of the TEI-based and therefore extremely complex schema via Excel is not an option for a possible automated transfer to TEI.

Nevertheless, the table structure can be prepared much more conveniently using an extremely simple XML document with a local namespace declaration. The cells of the header row – which corresponds to the row with the «label» role in XML – serve as reference cells for the mapped table structure. That is why a generalized table structure was imported as an XML source via the Developer tab into the Excel version of table 12, linking the generally structured table to the respective header row cells of table 12. After successful linking, the mapped table can be saved as XML via the Developer tab.

The XML tables exported in this way should now be relatively easy to code according to TEI. Tools such as Python – including looped RegEx-based searches and replacements – should support both the linking in Excel and the further processing in TEI to some extent.