Using TEI for textbook research

Lena-Luise Stahn
Georg Eckert Institute for
International Textbook Research
stahn@leibniz-gei.de

Steffen Hennicke
Georg Eckert Institute for
International Textbook Research
hennicke@leibniz-gei.de

Ernesto William De Luca
Georg Eckert Institute for
International Textbook Research
deluca@leibniz-gei.de

Abstract

The following paper describes the first steps in the development of an ontology for the textbook research discipline. The aim of the project WorldViews is to establish a digital edition focussing on views of the world depicted in textbooks. For this purpose an initial TEI profile has been formalised and tested as a use case to enable the semantical encoding of the resource 'textbook'. This profile shall provide a basic data model describing major facets of the textbook's structure relevant to historians.

1 Introduction

1.1 Textbook research and Digital Humanities

Textbook research has resulted in many forms of output, especially with its inter- and multidisciplinary approach as it is conducted at the GEI. The institute with its researchers from various different disciplines provides an excellent setting for cooperation and collaboration amongst these. As an increasing amount of projects is being executed in an interdisciplinary way and also in cooperation with the research infrastructure departments, the GEI evolves into a centre for the Digital Humanities, resulting in interdisciplinary and multimedial projects and digital plattforms. In recent years this development has proved as rather cumbersome for information retrieval and data reuse. The research results, most of them in digital form, end as legacy data in the sense of not being able to integrate them in the institute's information portals[1] because of data heterogeneity, therefore not beeing discovered and eventually not being used. This happens most often shortly after their production. The workflow usually consists of establishing a model for the digital representation each time anew. Rules or best practices on how to model textbook research data in a uniformal way in order to ensure their long term usability and stability do not exist. This is what makes the knowledge organisation situation at the GEI complicated.

1.2 Lack of Data Models and its impact on knowledge organisation in textbook research

But whereas models of the disciplines' knowledge exist in other fields, as for instance thesauri[2] or ontologies[3] providing access to the discipline's knowledge and information, the multidisciplinary character of textbook research aside its relatively short existence up to this day prohibit the development of a thorough data model, which would offer more powerful ways for knowledge organisation than controlled vocabularies used by the library for indexing purposes.

A large amount of the GEI's resources are available as XML-based fulltext after processing them with

[1] http://www.edumeres.net/

[2] e.g. AGROVOC Multilingual Agricultural Thesaurus, http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[3] e.g. Gene Ontology Consortium, http://geneontology.org/

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 181–186, Osaka, Japan, December 11-17 2016.

OCR (e.g. GEI-Digital[4], EurViews[5]). The Metadata Encoding and Transmission Standard[6] (METS) is used to describe them in a machine-readable way. Although METS is commonly used it does not offer adequate possibilities for the description of a textbook's content and characteristics, e.g. the specific school types and levels of education. By defining a model for structuring and encoding the data in a more thorough way, both data standardisation and integration on the one side and improved exploitation on the other side would be supported, improving the GEI data sources' retrieval, reuse and long term availability.

2 Project aims

2.1 Objective: Use Case WorldViews

A first attempt in this direction is the project WorldViews, funded by the BMBF and started in the beginning of 2015. A critical digital edition of textbook sources will be compiled, which is intended to serve historians with an entry point for discovering relevant, reliable and hard-to-find research materials on topics regarding textbooks. The materials may provide inspiration or even the corpus for medium scaled research endeavours or the foundation of more extensive research for additional sources. By establishing this digital collection comprising excerpts of textbooks and annotating them in regard to a particular research question (figure 1), a use case is being set up for developing and testing a first profile fit for the semantically contextualisation of the resources.

Annotation is interpretation, its purpose is to make a statement on the annotated element's appearance and/or meaning. By encoding elements in texts, aside from machine readability and long term availability, their detectability is ensured, in order to make them comparable and find relations. To support its semantic contextualisation the annotation language needs to be adapted to the specific text type, in the WorldViews context the textbook excerpts and editors' contributions. The use case WorldViews is the first project making an attempt to define such a specific annotation profile for the text type 'textbook'.

Since there existed no profile or standard for the description of textbooks, it was decided to use an adaptation of the TEI Guidelines[7] for encoding, mainly because of its common and well tested use for encoding text resources in the humanities, therefore ensuring compatibility and longterm usability in the context of Digital Humanities (DH) tools and methods. Furthermore it provided the flexibility and extension possibilities needed to adapt it to the project's research questions.

The Guidelines formulated by the TEI Consortium have become a quasi standard in the DH community for the encoding of historical text resources. Since their development in 1994 their XML-based structure has proved most useful for a broad range of text types, mainly based on its flexibility in adapting it to the respective discipline, resource types and research questions. In forming a TEI profile based on textbook research one of the project's expected results will be, whether TEI is able and an appropriate way to encode textbooks as well.

TEI commits to two essential axioms (TEI Simple Primer, 2016): First, a document is an "ordered hierarchy of content objects" (OHCO) (DeRose et al., 1990), and second, the presentation and the structure of a document can be cleanly separated. In the context of textual documents, both axioms are problematic. However, both axioms have proofed to be true often enough to be useful. Also the project team encountered these problems when trying to model each and every aspect of the text both on structural or presentation level at the same time. This most often resulted in a compromise between humanists and information scientists, receiving a model purposeful to both sides' present needs: as "an instance of the fundamental selectivity of any encoding. An encoding makes explicit only those textual features of importance to the encoder." (TEI Lite, 2012) Although deeply wished by the historians, most layout aspects were neglected.

[4] http://gei-digital.gei.de/viewer/
[5] http://www.eurviews.eu/nc/start.html
[6] http://www.loc.gov/standards/mets/
[7] http://www.tei-c.org/index.xml

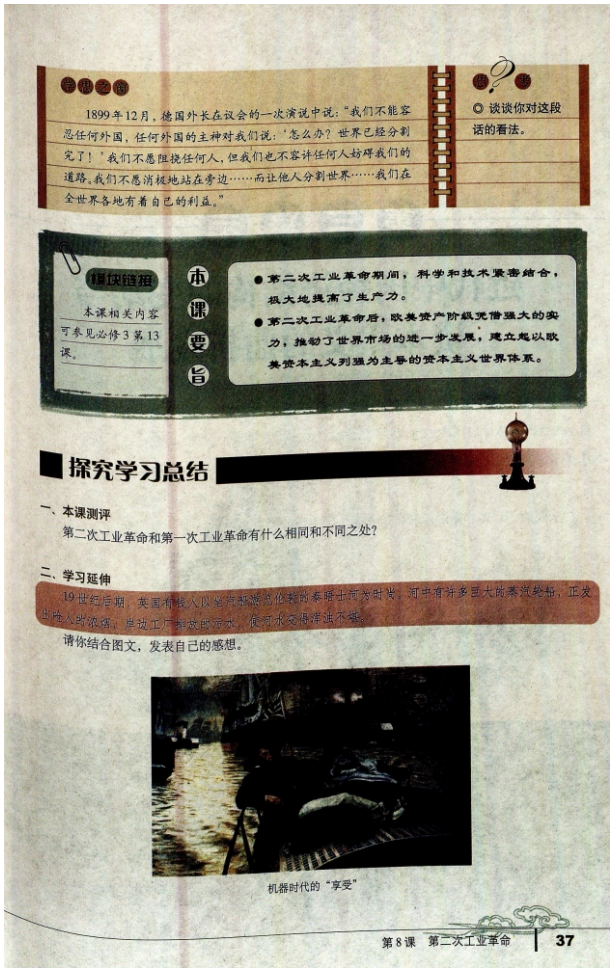

Figure 1: Example of a WorldViews source: chinese textbook excerpt

2.2 Approach: Using language technology in the humanities

First steps comprised of determining the characteristic elements of the used and established data within the project. This had to be done in close collaboration with the historians and cultural scientist involved in the project. Since they would eventually apply the data model, the usability of TEI during the research process could be tested as well when adopted by people not familiar with language technology tools.

Relevant questions had to be answered considering the kind of the arising research data. A necessity was to clearly formulate the annotations' purposes in order to avoid modeling data not relevant in the project context. This step showed the usual difficulties in asking the humanists of expressing their current - and possibly even future - research questions in an explicit way, fit for modeling them in a machine-readable form.

The wide variety of text types used in textbooks formed the problem of how to encode central characteristic text types, e.g. tasks ("pädagogische Anweisung") or clozes. It needed to be determinded how to treat visual elements like pictograms, infographics or timelines, often extended onto the next page, as well as image and text descriptions like image captions, texts in maps and marginalia, as they are increasingly used in modern educational resources. The humanists expressed their need especially for encoding these elements as they, as a major didactic method, play an important role in conveying the excerpt's world view.

A major issue applied to the corpus' multilingual character: the digital collection WorldViews, with its focus on world representations in textbooks from all around the world, would comprise of sources in various different languages, not necessarily written in latin alphabets or in a left-right direction (figure 2).

German label	English label	TEI element	TEI attributes / values
Quelle	source		
Logische Textgliederung	text division		@type=”chapter|section|part” @sample=”initial|final|medial| complete|unknown”
Textsegmente	Text segments	<seg>	@type=”authorText|assignment| question|definition|explanation| pedagogicalGuideline| pedagogicalIntroducation| chapterSummary|multipleChoice| dossier”
Überschrift	heading	<head>	@type=”[type of head]”
Absatz	paragraph	<p>	
Seitenumbruch	pagebreak	<pb>	@n=”[next page number]”
Zeilenumbruch	linebreak	<lb/>	
Tabelle	table	<table>	
Tabellenbeschriftung	table caption	-<head>	
Tabellenzeile	table row	-<row>	
Tabellenzelle	table cell	–<cell>	@role=”label” @cols=”[spaltenumfang]” @rows=”[zeilenumfang]”
Listen	lists	<list>	
Listenbeschriftung	list caption	-<head>	
Listenelement	list element	-<item>	
Zitate	quote	<q>	@type=”direct,indirect” @source=”[source of quote]”
Hervorhebung im Text	text highlighting	<hi>	@rend=”spacedOut|bold|italic| underline|strikethrough| blockCapitals|smallCapitals”
Anmerkung des Bearbeiters	editorial note	<note>	@type=”editorial”
Anmerkung des Schulbuchautors	note of author	<note>	@place=”foot|end|left|right” @n=”[Fussnotenzeichen]” @type=”footnote|endnote|gloss”
Verweis	reference	<ref>	@target=”[#xml:id]—[URI]” @type=”[type of reference]”
Grafisches Element	figure	<figure>	@type=”infographic|politicalMap| pictogram|photography|diagram| caricature|poster|painting|cartoon| speechBubble|arrows|drawing| mindmap|timeline” @place=”[indication of location on page]”
Titel der Abbildung	figure caption	-<head>	
Text zur Abbildung	figure text		
Beschreibung des Bildinhalts	figure description	-<figDesc>	
URI der Bilddatei	graphic URI	-<graphic url= />	@url=[image file path]
Bibliographische Angabe	bibliographic description	<bibl>	@xml:lang=de | en
Titel	title	-<title>	
Autor	author	-<author>	

Table 1: Extract of the TEI profile established for textbook resources in WorldViews

```
<body>

    <!-- page 36 -->
    <pb n="36"/>

    <div type="section">

      <head>世界市场的发展</head>

      <p>第二次工业革命比第一次工业革命的发展更为迅猛，也更为广泛。它在多个国家和几乎所有的工业领域同时展开，
促进了生产力的巨大增长，世界各地经济联系更加密切。</p>
      <p>第二次工业革命中出现的许多新型交通工具和通讯手段，大大加强了世界各地的联系。汽车越来越多，火车和轮船
越来越先进，交通运输日益便利；电报，电话的出现进一步加强了世界各地之间商业信息的交流与转播。</p>

      <!-- Bild auf Seite 36 -->
      <figure place="inline">
        <!-- [Bildlegende unten] -->
        <figDesc>铺设电缆</figDesc>
        <!-- [Bildlegende seitlich] -->
        <p>19世纪三四十年代，出现了有线电报。19世纪中后期，电报和电话把世界各地更紧密地联系在一起。1869年，从
英国伦敦到印度城市卡里卡特的电缆铺设完成，图为印度工人在铺设电缆。</p>
      </figure>

      <p>在第二次工业革命的推动下，世界市场进一步发展。1870年以后的三十多年间，世界贸易？增长了三倍左右。亚
洲，非洲和拉丁美洲等地区的非工业国家生产的粮食和原料源源不断地运往工业化国家，工业化国家生产的工业品则销往全世
界，国际分工日益明显。</p>
      <p>第二次工业革命期间，资本主义列强在全世界划分殖民地和努力范围，掀起了瓜分世界的狂潮。19世纪末20世纪
初，世界基本被资本主义列强瓜分完毕，亚洲，非洲和拉丁美洲广大地区基本上都沦为殖民地或半殖民地。资本主义国家在输
出商品，掠夺原材料的同时，直接向殖民地或半殖民地输出资本；殖民地和半殖民地的民族资本主义工业开始了艰难的发展历
```

Figure 2: Example of the above chinese textbook excerpt's fulltext in TEI

Some examples of the decisions for TEI elements to be used on the type 'textbook':

- <figDesc>is important because images on a textbook page may not be licenced yet and are therefore not displayed in the frontend. In such cases, <figDesc>allows to provide a meaningful textual description of the images and of those aspects of the image that are relevant to the narrative.

- @type:
  In order to retain flexibility regarding future extensions of the profile, text passages have been qualified by means of custom data values for type attributes. That allows for easy complementing of new relevant types of text passages.

- marginalia (defined as note of the textbook author):
  <note type="gloss" place="outer margin">Ausbildung des Ritterstandes</note>with possible types: @type="footnote—endnote—gloss"

3 Knowledge Organisation: From specific to general data modeling

3.1 Expected results

The project's major outcome is expected to be a profile for text resources meeting exactly the purpose of encoding elementary characteristics of textbooks. By the project's process of closely collaborating with the historians the profile's easy handling and managebility by people not used to working with language technology tools will be ensured. By referring to common standards also the compatibility with existing profiles and guidelines like TEI-Simple (TEI Simple Primer, 2016) and TEI-Lite (TEI Lite, 2012) is considered in order to retain its long term usability. These aspects will suppport the WorldViews profile to become the first version of a basic format for textbook sources.

The project will have impact on the GEI's research working processes as well: a workflow is formalised and tested which will support future interdisciplinary projects to determine their major elements based on the research output the projects are expected to generate. It will serve as an example of what major research questions could be, how difficult text types and layouts can be handled. Furthermore it will show how information and computer scientists and humanists can communicate on a mutual level in order to achieve the needed data model formulated in a way meeting both sides of researchers' needs as exactly as possible.

3.2 Future Work

Future work may address the question of how the use case WorldViews can serve as a first survey, forming the basis for a general data model, outlining the research done at the GEI. This data model, integrated in the institute's information retrieval tools and portals, would form a major part in supporting the knowledge organisation at the GEI. The WorldViews model's ability to support the long-term objective of formalising an ontology for the textbook research needs further research and use cases. This development could eventually support the GEI's intention of becoming an internationally acting centre in this research field.

References

DeRose, Steven, David Durand, Elli Mylonas, and Allen Renear. 1990. *What is text, really?* Journal of Computing in Higher Education, 1 (2): 3-26.

TEI Simple Primer, `https://github.com/TEIC/TEI-Simple`.

TEI Lite, `http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_lite.doc.html`.