Curating textbooks – What are humanities scholars looking for and how can we support their research?

Maret Keller

Georg-Eckert-Institute – Leibniz-Institute for International Textbook Research (GEI), Braunschweig, Germany
keller@gei.de

Abstract. As textbooks both describe and help shape societies, they represent a valuable source for a wide range of humanities research questions and methods. This paper is concerned with the variety of information requirements researchers from the humanities may have when working with these sources. It describes the efforts made at the Georg-Eckert-Institute for International Textbook Research to make the data contained in digitised textbooks as accessible as possible and presents some considerations on the usefulness of more thorough curation and the use of digital humanities tools.

Keywords: digital curation, textbook research, information requirements.

1. Introduction: digital textbooks at the GEI

Textbooks are mass media present in every classroom and used by every school child. History and geography textbooks in particular, viewed down the ages, show us what children and adolescents of various times and cultures were able and expected to learn and "know" about the world. This makes them highly significant cultural artefacts that can teach us just what different societies at different times considered to be "true", and important enough to be taught to the younger generations and how they went about teaching these things.

The research library at the Georg-Eckert-Institute for International Textbook Research (GEI) hosts some 175.000 textbooks and curricula from 160 countries. In addition to descriptive and subject cataloguing, the GEI aims to provide digital accessibility of data relevant to textbook research. The digital textbook library GEI-Digital is accessible online and has recently reached the milestone of the 5000 books/1 million pages of historic German textbooks in the public domain.[1] Other projects such as "EurViews" digitise fragments of more recent books from various countries and publish them as part of a digital scholarly edition.[2] The Edumeres portal offers services such

[1] http://gei-digital.gei.de/viewer/
[2] http://www.eurviews.eu

as a worldwide database on textbook systems, OA publications, reviews and digitised curricula to the research community and provides access to a range of digital projects.[3]

The primary objectives of data curation at the GEI are discoverability, sustainability and reusability. A closer look at the GEI-Digital project reveals both the achievements and the challenges of the Institute's work in this field to date: There is metadata, documentation, facsimiles and text OCRed with ABBYY finereader, persistent identifiers (URN/PID), and presentation with DFG and Intranda viewers. The data is online, CC 1.0 and downloadable via OAI interface. The Goobi workflow involves some very basic structuring which provides a basis for navigation within a book by chapters (see fig.1), and for facet browsing for such content as images, advertisements and forewords. The textbooks are grouped by subject and epoch and the metadata enriched by the librarians allows for further facet browsing. The fulltext search facility is limited due to flaws in the OCR, above all in highly structured and/or gothic script texts in which large numbers of pre- and early-twentieth-century German-language books were printed (see fig. 2).

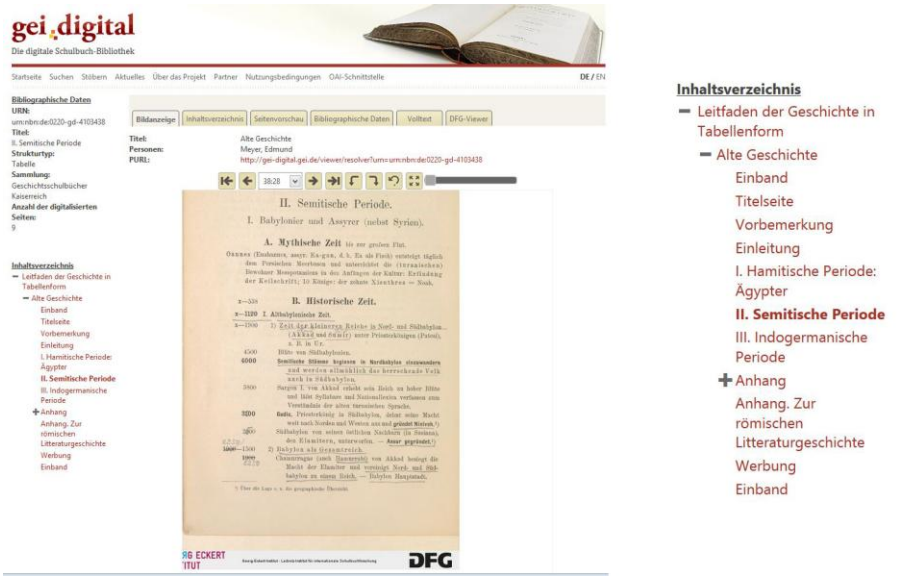
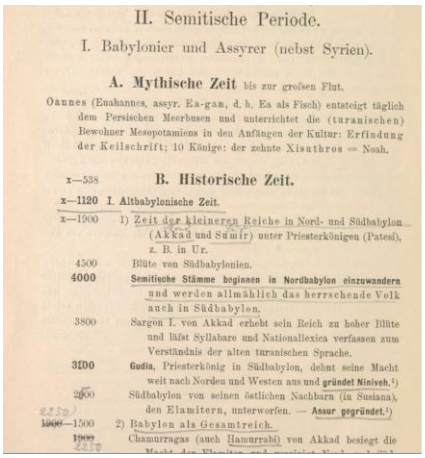


Fig. 1. Example of GEI-Digital interface; detail of navigation (with sample from Meyer, *Leitfaden der Geschichte in Tabellenform*, p.28).[4]

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

[4] Meyer, Edmund, *Leitfaden der Geschichte in Tabellenform: Alte Geschichte*, Berlin: Weidmann 1890, p. 28, in: GEI-Digital, http://gei-digital.gei.de/viewer/resolver?urn=urn:nbn:de:0220-gd-4103438, CC1.0.



II. Semitische Periode.

I. Babylonier und Assyrer (nebst Syrien).

A. Mythische Zeit bis zur großen Flut.

Oannes (Euahannes, assyr. Ea-gan, d. h. Ea als Fisch) entsteigt täglich dem Persischen Meerbusen und unterrichtet die (turanischen) Bewohner Mesopotamiens in den Anfängen der Kultur: Erfindung der Keilschrift; 10 Könige: der zehnte Xisuthros = Noah.

x—538 **B. Historische Zeit.**

x—1120 I. Altbabylonische Zeit.

x—1900 1) Zeit der kleineren Reiche in Nord- und Südbabylon (Akkad und Sumir) unter Priesterkönigen (Patesi), z. B. in Ur.

4500 Blüte von Südbabylonien.

4000 **Semitische Stämme beginnen in Nordbabylon einzuwandern und werden allmählich das herrschende Volk auch in Südbabylon.**

3800 Sargon I. von Akkad erhebt sein Reich zu hoher Blüte und läßt Syllabare und Nationallexica verfassen zum Verständnis der alten turanischen Sprache.

3100 **Gudia**, Priesterkönig in Südbabylon, dehnt seine Macht weit nach Norden und Westen aus und **gründet Niniveh.**[1])

2000 Südbabylon von seinen östlichen Nachbarn (in Susiana), den Elamitern, unterworfen. — **Assur gegründet.**[1])

~~1000~~—1500 2) Babylon als Gesamtreich.

~~1900~~ Chamurragas (auch Hamurrabi) von Akkad besiegt die

38:28

 II. Semitische Periode.
 I. Babylonier und Assyrer (nebst Syrien).
 -A-. My tlllSChe Zeit bis zur grofsen Flut.
 Oannes (Eualiannes, assyr. Ea-gan, d. li. Ea als Fisch) entsteigt täglich
 dem Persischen Meerbusen und unterrichtet die (turanischen)
 Bewohner Mesopotamiens in den Anfängen der Kultur: Erfindung
 der Keilschrift; 10 Könige: der zehnte Xisuthros = Noah.
 x—538 B. Historische Zeit.
 x—1120 I. Altbabylonische Zeit.
 x 1900 1) Zeit der kleineren Reiche in Nord- und Südbabylon
 (Akkad und Sumir) unter Priesterkönigen (Patesi),
 z. B. in Ur.
 4500 Blüte von Südbabylonien.
 4000 Semitische Stämme beginnen in Nordbabyion einzuwandern
 und werden allmählich das herrschende Volk
 auch in Südbabylon.
 3800 Sargon I. von Akkad erhebt sein Reich zu hoher Blüte
 und läfst Syllabare und Nationallexica verfassen zum
 Verständnis der alten turanischen Sprache.
 3100 Gudia, Priesterkönig in Südbabylon, dehnt seine Macht
 weit nach Norden und Westen aus und gründet Niniveh.1)
 2000 Südbabylon von seinen östlichen Nachbarn ~(in Susiana),
 ZZSvJ den Elamitern, unterworfen. — Assur gegründet.1)
 IdOÖ—1500 2) Babylon als Gesamtreich.
 1306- Chamurragas (auch Hamurrabi) von Akkad besiegt die
 Macht der Elamiter und vereinigt Nord- und Süd¬
 babylon zu einem Reich. — Babylon Hauptstadt.
 ') Über die Lage s. u. die geographische Übersicht.

Fig. 2. Facsimile and OCR (Meyer, *Leitfaden der Geschichte in Tabellenform*, p.28).[5]

As usual, with limited resources for infrastructure development, decisions about priorities have to be made. For GEI-Digital the priority is offering large quantities of digitised material. What kind of further curation would be possible and desirable? What do we do with a million pages? Over the past two-and-a-half years, some of these possibilities, like further TEI-XML Annotations and Topic Modeling, have been tested in three Digital Humanities projects run at the GEI.

2. What do humanities researchers search for?

2.1 Searching for texts

Humanities researchers search for texts, to find evidence in relation to their specific research questions or leads for tentative ideas. They may need to create a corpus, or may choose to work with a given one.

In most cases, the identification of, access to and collation of source materials constitutes a substantial part of the research itself, as the texts needed come from different domains and contexts and are to be found in a wide range of libraries, archives and databases, in various media and formats. Research of this kind might look for a subset of the textbooks contained in GEI-Digital, for the purpose of analyzing them along with other resources[6].

Other research may call for a comprehensive ready-made corpus; examples might be editions of transcribed manuscripts on a certain topic or by specific authors, or datasets, such as a chat corpus, a newspaper archive or the textbooks in GEI-Digital.

---

5 See note 4.

6 For examples of research conducted with textbooks, see the open access publications in the GEI's DSpace Repository: http://repository.gei.de/handle/11428/2.

2.2 Searching within texts

Scholars might be looking for non-text materials within texts, such as pictures and tables, either for their own sake or for studying their purpose and functionality within the text; they might be looking for information about events, for historical facts in chronicles, newspapers, or minutes. They might be looking for potential knowledge or even (what they assume to be) absolute truths contained within scientific prose or religious texts. They may be looking for things that grab their attention, move and entertain them because they are interested in how a text works within the mind of a reader. They will often be looking for evidence within a text of phenomena residing outside the text: perhaps the mental state of the author(s), the influence of specific ideas, the worldviews of their generation, class or gender, changes in language use and historical discourse in general. They might even be interested in what is not there, if they are interested in gaps of knowledge or taboos.

2.3 Searching for contexts

It is a conviction and proposition of humanities that semantics are always dependant on context, never absolute. Even if all the source material needed is offered in a service such as GEI-Digital, the researcher will still need additional information from a variety of domains. Take, for instance, the most common contextual relationship, that between a text and its author. Knowing his or her name helps us to find a text by a certain author; getting to know the text, we can make assumptions about the author's motivations for writing it, or about his or her character. A text and the circumstances of its production and publication (i.e. the context) can be seen as metadata of each other. Even two different texts can become metadata of each other.
Researchers also search for similarities and contrasts. Comparing texts (or corpora) and looking for (textual) contexts may lead to the reconstruction of processes such as the dissemination of ideas via intertextual relations, to discoveries around changes in language, attitudes and concepts, or indeed to the uncovering of plagiarism.

3. How can we help?

Digital infrastructure for textbook research has to cater for research interests from a wide range of disciplines and perspectives. The specific information needed may include materials and media within the data, (historical) facts, style, (higher) truths, gaps, and examples of (linguistic) discourse on a potentially unlimited number of topics. At the same time, the texts in question have various characteristics to be considered: mixtures of authorial text and quoted source materials in history textbooks, mixtures of genres in reading primers, the existence or otherwise of self-referential introductions, footnotes, remarks on educational methodology directed to the teacher, and assignments for the students. All this considered, there are essentially three straightforward aims of the digital curation of textbooks:

To aid researchers in *searching for* text and *searching within* texts, we need to develop:

1. Good metadata. Information is needed about education systems and about the demographic make-up of school students, such as age, gender, and religion. This might entail the use of well-established metadata models, such as CLARIN's CMDI metadata, for visibility and compatibility.
2. Smart information retrieval. The "TextbookCat" project enables browsing by country, level of education, school subject and the local classification used by the research library at the GEI.[7] Development of a search function for all the GEI data from its range of projects is in progress.
3. Good data. Corrected OCR is needed to permit reliable and thorough searches. This represents a challenge as double keying is out of the question for such a large corpus.

To aid the *search for contexts*, documentation is needed about the nature of digital corpora. Other services relevant to textbook research include the collection of academic literature held by the GEI's research library and the databases and services of the Edumeres portal. Compatibility with services such as CLARIN-D[8] and the German Text Archive (DTA)[9] will increase visibility and allow the use of their search functions for reference material and of tools for further linguistic analysis.

4. What about DH-tools?

4.1 TEI-XML

Annotations in textbooks can help researchers find specific features of textbooks and even make texts "compatible" with other texts and corpora. In a CLARIN- curation project[10], we added annotation in accordance with the basic format of the DTA[11] after manually correcting the OCR. This was great for layout and the use of the DTA's linguistic search tools and ensured compatibility with a large reference corpus. We found that specific information, such as the aforementioned assignments for students, could not be coded; the "WorldViews" project is working on this currently by developing a TEI profile for textbooks.[12]
But annotations represent interpretations of a text, and in all but the most basic cases, this is part of the humanities scholar's analysis. Thorough documentation is of great importance: Fictitious persons such as Goethe's Werther or Donald Duck might not be tagged (and thus searchable) as a named entity. Expressions such as "home",

7 http://tbcat.edumeres.net/
8 http://de.clarin.eu/de/; https://vlo.clarin.eu/
9 http://www.deutschestextarchiv.de/
10 http://www.clarin-d.de/en/curation-project-9-1-mordern-history
11 http://www.deutschestextarchiv.de/doku/basisformat
12 http://www.gei.de/abteilungen/europa/bruchlinien/worldviews-die-welt-im-schulbuch.html

"middle east" and "paradise" might be highly relevant for some researchers but will most probably not be tagged as geographical entities.

4.2 Ontologies

Like any DH-tool, ontologies have to be subjected to a "tool critique", to ascertain whether they (and their configurations) are really adequate to use. An problem that frequently crops up is related to historic semantics: The "knowledge" of a 1900s natural history textbook, for example, might be interesting to transcribe into an ontology. But you would not be able (and would not want) to use this ontology when working with a 2016 textbook a similar subject.

4.3 Topic Models

If we use topic models merely to find out what topics might be present, the algorithm as a tool may remain a black box, an oracle, that will hopefully provide hints about what to look for manually. If its outcomes are to serve as authoritative evidence, then the researcher has to understand (and be able to explain) how it works. Can we achieve valid results using dirty OCR? How does the preprocessing influence the results? The "Children and their World" project evaluates these questions.[13]
For example, textbooks may contain a great many textual genres, by different authors, that obey their own rules and may be included in the collection for many different purposes (to represent a literary style, an epoch, to give a moral or historical information, for instance. The humanities researcher has to deal with the fact that these characteristics and contexts may not be distinguished by an algorithm. Being used to look for the outstanding and singular, he or she needs to learn what is to be learned from stochastics and statistics.

5. Conclusions:

In order to provide international textbook research with well-curated data, we have to bear in mind the different information needs of researchers from different domains. The importance of precise and detailed metadata with adequate information retrieval is as undisputed as the utility of transcription true to the original. The usefulness of DH methods such as XML annotations, ontologies and topic models has to be discussed and investigated further, weighing it against the time and effort required for their implementation. Information scientists and humanities researchers have to figure out how to overcome – or bring together – specific and generic, unique and pattern orientated, ambiguity friendly and distinctness bound interests and methods.

[13] http://www.welt-der-kinder.gei.de/en