# Fusing International Textbook Collections for Textbook Research


**Christian Scheel and Ernesto William De Luca**

**Abstract** The investigation of sanctioned knowledge for the formation of the young generation is a subject of textbook research. Additionally, textbooks are gaining importance for historical research in other disciplines, in the search for "popular knowledge", as they reflect worldviews, thought flows and desired knowledge. Therefore, it is important to have knowledge of textbooks from around the world. This work describes the fusion of international textbook collections and the usage of classification schemes, covering textbook specific characteristics. The resulting research tool for international textbook research is a prototype for fusing any other metadata-based web databases to raise research from national to international level.

**Keywords** Textbook research · Digital archives · Fusing web databases

## 1 Introduction

Although the field of textbook research is relatively new, there are many independent textbook collections available dating back to the year 1979, covering textbooks from the 18th to the 21st century [1]. Those researchers who were interested in investigating knowledge and culture found in textbooks created their own archives or databases. Besides containing textbooks and because of different areas of interest, these collections are very diverse in size and quality of information. Some collections are physical available and others are just bibliographies. These textbook researchers have not been librarians; they often did not follow any standards or created their own standards, but the collections have in common that the textbook specific characteristics have been documented very well. Such information as "country of use", "school subject", "type of school", or "level of education" is crucial for textbook research.

C. Scheel · E. W. De Luca (✉)
Georg Eckert Institute for International Textbook Research, Celler Str. 3,
38114 Brunswick, Germany
e-mail: deluca@leibniz-gei.de

C. Scheel
e-mail: scheel@leibniz-gei.de



H. Kremers (ed.), *Digital Cultural Heritage*,
https://doi.org/10.1007/978-3-030-15200-0_7

Usually, no library is annotating textbooks at this level of detail, because textbooks are treated like any other book. Hence, to support international textbook research, these textbook collections had to be analyzed, cleaned, unified and made available to the public.

As starting point to unify these collections, we have chosen the international textbook collection, which is physically available at the research library of the Georg Eckert Institute, Brunswick, Germany. For describing 180,000 textbooks, the librarians created a classification scheme for textbook specific characteristics. Unfortunately, this classification was partly focused on the national level, so that "type of school" was only annotated for German textbooks and "periods" covered only periods in time that have been important to German history. In addition, "federal state of use" was only used for textbooks where "country of use" was Germany. The research, the textbook collection and hence the local classification scheme of the Georg Eckert Institute are primarily focused on educational sciences, history, geography, political science and religious sciences [2–4], where other collections also include math or even cooking.

In this work, we present results of the analyzation of international textbook collections and generalize found problems. We go on, proposing how to create a classification scheme covering all textbook specific characteristics on international level with as much reuse of existing schemes and standards. We show how information in textbook collections are mapped to codes of the new classification scheme and how these codes are the foundation of a language independent research tool for textbook research.

## 2 Textbook Collections

To guarantee that the young generation will follow the rules and views of their leaders, knowledge is often sanctioned. This happened in the past and certainly happens now. Besides other areas of research, textbook researchers are analyzing the worldviews, thought flows and desired knowledge found in textbooks, providing additional information to questions like "How and why have historical events been supported by the people?" The youth is not questioning the knowledge in textbooks, which makes them particularly vulnerable. By comparing textbooks over time and internationally, sanctioned knowledge can be revealed. Therefore, textbook researchers started collecting textbooks in their area of interest.

International textbook collections have emerged mostly as an initiative of individual researchers (groups), and from the need to catalog textbooks because there was no explicit indexing of textbooks by the libraries. Unlike the Georg Eckert Institute, where all textbooks are also part of the research library, the databases of international textbook researchers are usually only bibliographies. However, this is not a disadvantage, otherwise information about textbooks, e.g. when books are deposited in monastic libraries, are simply unknown. In addition, the quality of the information can be regarded as high, especially for textbook specific characteristics.

Besides the textbook collection of the research library of the Georg Eckert Institute, three other collections are referenced in this work: EDISCO, MANES and the Emmanuelle Textbook project. Analyzing and comparing these collections have played a major part for fusing these databases. The University of Turin's Research Center for Digitization and Creation of Digital Libraries for the Humanities hosts the EDISCO database, which has around 30,000 records of Italian textbooks from the 19th and 20th centuries. The Department of History of Education and Comparative Education hosts the MANES database, which has approximately 35,000 records of Spanish, Portuguese and Latin American textbooks from the 19th and 20th centuries. The metadata in MANES is recorded according to its cataloging guidelines [7]. In 1979, Alain Choppin, a pioneer of textbook research, started to create the "Emmanuelle Textbook Project". This collection includes French textbooks dating back to 1789 [8, 9].

## 3 Fusing Web Databases

Before the International TextbookCat was implemented, a great part of textbook research was networking. Researchers needed to know each other personally to gain access to textbook collections from other countries. However, for researchers interested in the field of textbook analysis, finding textbook collections was not even the biggest hurdle. Because textbook researchers usually have not been librarians, textbook collections have not been standardized, resulting in frustrating observations. In the following, some of these observations are listed:

- Although books are described, common fields for describing books are missing.
- For accessing the data, logins are needed.
- Accessing the data is only possible within the corresponding institute.
- The user interface is not intuitive and needs training or supervision.
- The user interface is monolingual.
- The metadata is monolingual.
- The metadata is not limited and contains spelling errors.
- The data can not be retrieved via interfaces.

Hence, even when having access to another textbook collection, the results have not been comparable. The problems can be divided into three fields: access, user interface and data. While the whole idea of fusing these web databases is about unrestricted access and a common user interface, there is still the need to retrieve and fuse the collections automatically. Therefore, interfaces and standards are needed.

### 3.1 The Need for Access

To break the existing structures, we encourage those who maintain a textbook collection, to also offer open access to everyone. Hence, when starting to fuse these web databases, we made clear that anybody joining the project would have to migrate to a common library system. Library systems not only provide access for users, but also enable automated data collection through the standardized OAI-PMH interface [5].

### 3.2 The Need for Standards

Before fusing international web databases, there is the need for common classification schemes. While all characteristics of books should be covered by the used library system. All other metadata must be represented in a consistent format or at least must be mappable into such presentation.

The textbook collection of the research library of the Georg Eckert Institute contains over 180,000 textbooks. Although about 100,000 of these textbooks are written in German, the collection contains textbooks from 181 countries. To better describe these textbooks, the librarians developed a local notation, a classification system that could be used to describe every book in the collection in detail (see Table 1 ↓). The "country of usage" is a textbook specific feature, because the country where a book was printed is not necessarily the country where the book is used. Hence, country of usage is very important to estimate the impact of a textbook. In this collection, the class "state" contains only codes for the German states. Due to the limited research focus, only subjects such as history, social studies, geography and religion, as well as language lessons, can be found in "school subject". Textbooks for mathematics are therefore not part of the collection (and the local notation), as their significance for the historical and cultural studies of (international) textbook research is considered to be comparatively low. The class "level of education" corresponds almost exclusively to the classification of the International Standard Classification of Education (ISCED) of UNESCO [6]. The notation for "school type" only covers the German school system, even if approximately 100,000 textbooks in this collection have not Germany as "country of usage". The Germany-centric view is continued in the class "period". This class is redundant to the publication year, but allows a classification according to 15 historically relevant and Germany related epochs. The textbook specific class "document type" distinguishes between textbook, syllabus, teaching material, teacher's manual, collection of exercises, etc. Therefore, it may not be confused with the general type, covering book, e-book or video.

Ideally, the collections of the international partners could have been one-to-one mapped to the data structure of the Georg Eckert Institute. That this could not be the case was foreseeable in advance, since the school systems differ from one country to another and thus there is often no corresponding code in this classification scheme. Further reasons have been revealed when analyzing the international collections.

**Table 1** Local notation for textbook specific characteristics of the Georg Eckert Institute, where _ represents the placeholder for numbers or letters

| Code(s) | Class | Expressions |
|---|---|---|
| l_ _ _ | Country (of usage) | 181 |
| b_ _ / b_bz | State (Germany) | 16/4 |
| u_ _ _ | School subject | 15 |
| k_ _ | Level of education | 7 |
| s_ _ | School type (Germany) | 11 |
| z_ _ _ | Period | 15 |
| d_ _ | Document type | 12 |

**Individual fields in the collections**. Since the creators of the collections were not librarians, in some cases only those characteristics were recorded that are important for textbooks in particular, but not books in general. For example, titles, subtitles, and title additions are recorded separately according to library standards, whereas a "layman" would collect that information as a whole.

**Interpretation of individual fields**. A further challenge were fields whose names were ambiguous, so that they have been interpreted differently by different people. For example, the field "usage" was interpreted by MANES creators both as "usage for" and as "usage as", even though there are cataloging guidelines [7]. Hence, this field contained individuals and groups, as well as document types.

**Differences in the degree of detail**. When looking at the collections, the disadvantage of free text fields could be seen. The absence of mandatory expressions for certain attributes, for example, was reflected in 155 different expressions for school subjects in EDISCO and 85 school subjects in MANES, which contrasted with the 15 school subjects of the GEI classification system. Even though a large part of the variants was due to spelling errors, after the adjustment, 86 new school subjects remained, which had to be taken into account in the international classification system. This made a fundamental decision necessary on whether to generalize properties (fewer options), specialize (all options) or simply ignore non-mappable properties. Generalizing or ignoring would lead to the loss of information, while specializing has the disadvantage that one has to (unambiguously) translate each expression (including future ones) into each supported language.

**Different understanding of matter**. Can cookbooks be seen as textbooks? Should math books be considered, even when there is no use of application in textbook research?

**Entries are difficult to map**. Due to different education systems in different countries, fields such as the "school type" can not be mapped. In the research tool, filtering on such a field would usually only lead to results from a specific country, which is not wanted.

### 3.3 The Need for Multi-lingual User Interfaces

In the past, there were many limiting factors for textbooks as a source of the humanities. One of them are the languages. When working with a textbook collection in another language, the researcher had to have at least a basic understanding of that language in order to navigate through this catalogue and additional effort was needed to translate the metadata by hand. When fusing international catalogues, there is the need to present metadata in the researcher's mother tongue.

## 4 Fusing Metadata

The task of fusing web databases can be broken down to the task of fusing metadata. Each possible expression for any feature (like "school subject", "level of education", etc.) coming from any web database needs to be annotated language independently.

### 4.1 Classification Schemes and Norms

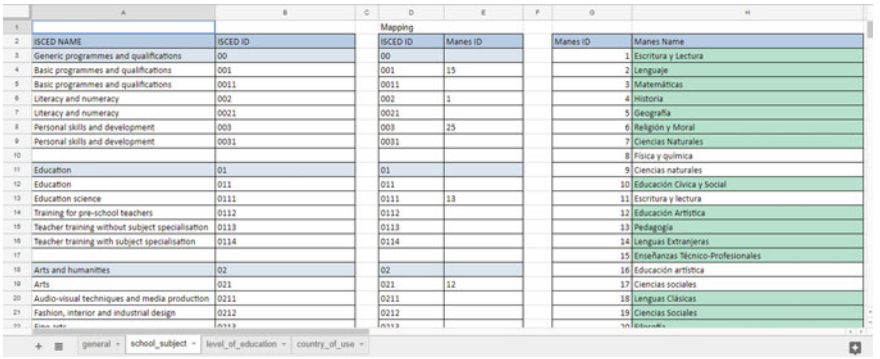
When "country of usage" is "Germany" in an English database and "Deutschland" in a German database, then both expression have to be mapped to the same code. While there are established norms for representing countries (ISO 3166-1), there might be missing norms for individual features like "school type". The first task for fusing metadata is too search for classification schemes which can be reused.

One of the most convenient detail about classification schemes and standards is the usage of codes. When representing "Germany" with "de", this code can easily be mapped back to "Germany" or "Allemagne", if French is you mother tongue. Hence, the need for multi-lingual interfaces is actually satisfied on the way.

There often are no norms for representing individual features. For instance, there are no codes for representing cities. In such cases, we are using the English translation as code and collected the necessary translations from knowledge sources like DBpedia [10] (see owl:sameAs relation[1]).

In the field of textbooks, we were happy to find standards for "school subject" and "level of education". The International Standard Classification of Education (ISCED [6]) provides hierarchical schemes for both features. Hence, we only had the task of translating all codes into other languages. When applying these hierarchical schemes, another convenient detail was revealed. While there are 120 codes for representing school subjects, there are only 10 top-level codes (like "Fine arts"or "Natural sciences, mathematics and statistics"). The most specific codes can be found in the third level. Therefore, there was no need to decide whether we want to represent a feature specifically or generally. The specific expression "Math" is implicitly part of

[1] http://de.dbpedia.org/page/Berlin.



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | Mapping | | | | |
| 2 | ISCED NAME | ISCED ID | | ISCED ID | Manes ID | | Manes ID | Manes Name |
| 3 | Generic programmes and qualifications | 00 | | 00 | | | 1 | Escritura y Lectura |
| 4 | Basic programmes and qualifications | 001 | | 001 | 15 | | 2 | Lenguaje |
| 5 | Basic programmes and qualifications | 0011 | | 0011 | | | 3 | Matemáticas |
| 6 | Literacy and numeracy | 002 | | 002 | 1 | | 4 | Historia |
| 7 | Literacy and numeracy | 0021 | | 0021 | | | 5 | Geografía |
| 8 | Personal skills and development | 003 | | 003 | 25 | | 6 | Religión y Moral |
| 9 | Personal skills and development | 0031 | | 0031 | | | 7 | Ciencias Naturales |
| 10 | | | | | | | 8 | Física y química |
| 11 | Education | 01 | | 01 | | | 9 | Ciencias naturales |
| 12 | Education | 011 | | 011 | | | 10 | Educación Cívica y Social |
| 13 | Education science | 0111 | | 0111 | 13 | | 11 | Escritura y lectura |
| 14 | Training for pre-school teachers | 0112 | | 0112 | | | 12 | Educación Artística |
| 15 | Teacher training without subject specialisation | 0113 | | 0113 | | | 13 | Pedagogía |
| 16 | Teacher training with subject specialisation | 0114 | | 0114 | | | 14 | Lenguas Extranjeras |
| 17 | | | | | | | 15 | Enseñanzas Técnico-Profesionales |
| 18 | Arts and humanities | 02 | | 02 | | | 16 | Educación artística |
| 19 | Arts | 021 | | 021 | 12 | | 17 | Ciencias sociales |
| 20 | Audio-visual techniques and media production | 0211 | | 0211 | | | 18 | Lenguas Clásicas |
| 21 | Fashion, interior and industrial design | 0212 | | 0212 | | | 19 | Ciencias Sociales |
| 22 | Fine arts | 0213 | | 0213 | | | 20 | Filosofía |

general | school_subject | level_of_education | country_of_use

**Fig. 1** Example for mapping Spanish data into the ISCED classification for school subjects. On the left, the target classification can be seen. On the right, all found expressions for school subjects (including spelling mistakes) are listed. The mapping is defined in the middle of this spreadsheet

"Mathematics and statistics" and the most general representation "Natural sciences, mathematics and statistics".

## *4.2 Mapping*

The most time consuming part of fusing metadata is defining the mapping for each web database. An expert is needed to link each expression from the data to one or more codes in the classification scheme. The best expert is always the owner of the data, because this person knows the real meaning of expressions and will not make assumptions based on translations.

We experienced, that using spreadsheets (like Google Spreadsheets; see Fig. 1) for defining the mapping was a good choice, because of the intuitive nature of tables.

After retrieving a dataset (through OAI-PMH interface [5]), we automatically collected all expressions for each feature, gave each expression a unique identifier, created the spreadsheet and asked to link these identifiers to the codes of the respective classification scheme. These files can be revised at any time, because the tools do not manipulate the original data, but the representations in the search index.

## *4.3 The Used Tools*

Besides having developed tools for creating, filling and reading the spreadsheets automatically and observing all databased for new or changed entries, most of the tools used are open source tools.

For reading the textbook collections, applying the mapping and writing the results to a search index, the tool MARC4J[2] is used. The search index, which is the only storage where the fused data is kept, is part of a Solr[3] server.

The user interface is based on VuFind.[4] Besides being already translated for several languages, this discovery system supports saving searches or bookmarking books in individual list. This has shown to be very helpful for managing books for specific research questions.

## 5 The International TextbookCat

In the field of textbook research, there have been many textbook collections, which were hard to access. First, it was difficult to find these collections, and secondly, the textbooks were usually annotated in the language in which they have been written.

For creating the International TextbookCat,[5] a research tool, which covers international textbook collections, these catalogues had to be analyzed and fused on metadata level. The most important part was to find classification schemes for each feature (like "school subject") and to define mapping rules to map the expressions found in the catalogues into these schemes.

In the resulting research tool, the classification schemes are reflected as facets. These facets allow to search for textbooks that have certain attributes, independent from the language each attribute was formulated in its original database. Representing attributes as codes also helped to present every metadata in the researcher's mother tongue.

The problems to be solved before fusing textbook collections are also found in similar areas. In any research field, there are historically grown collections (of any items) which are important to the field, but can not be accessed properly. The process of analyzing, refining and mapping can surly be applied to any of these collections, while also helping to preserve them.

## References

1. Ossenbach, G.: „Textbook databases and their contribution to international research on the history of school culture", History of Education & Children's Literature, pp. 163–174, 2014.
2. Fiedler, M., Scheel, C., Weiß, A., De Luca, E.: „Welt der Kinder. Semantisches Information Retrieval als Zugang zu Wissensbeständen des 19. Jahrhunderts", FVWG und GCDH Workshop für Wissenschaftsgeschichte und Digital Humanities in Forschung und Lehre, 2016.

[2] https://github.com/marc4j/marc4j.

[3] http://lucene.apache.org/solr/.

[4] https://vufind.org.

[5] http://itbc.gei.de.

3. Fuchs, E., Kahlert, J., Sandfuchs, U.: Schulbuch konkret: Kontexte - Produktion - Unterricht. Klinkhardt, 2010.
4. Fuchs, E., Niehaus, I., Stoletzki, A.: Das Schulbuch in der Forschung: Analysen und Empfehlungen für die Bildungspraxis. V&R Unipress, 2014. URL http://www.gei.de/publikationen/eckert-expertise/.
5. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2015-01-08. URL http://www.openarchives.org/OAI/openarchivesprotocol.html.
6. Silke L. Schneider, (2013), The International Standard Classification of Education 2011, in Gunn Elisabeth Birkelund (ed.) Class and Stratification Analysis (Comparative Social Research, Volume 30) Emerald Group Publishing Limited, pp. 365–379.
7. Gabriela Ossenbach: „Research about school handbooks in Latin America: The project MANES contribution", Historia de la Educación, 2013. URL http://revistas.usal.es/index.php/0212-0267/article/view/10797.
8. Alain Choppin: „EMMANUELLE: a data base for textbooks' history in Europe", Historical Social Research, pp. 52–58, 1989. URL http://nbn-resolving.de/urn:nbn:de:0168-ssoar-51666.
9. Alain Choppin: „The Emmanuelle Textbook Project", Journal of Curriculum Studies, pp. 345–356, 1992. URL http://dx.doi.org/10.1080/0022027920240404.
10. Auer, Sören & Bizer, Christian & Kobilarov, Georgi & Lehmann, Jens & Cyganiak, Richard & Ives, Zachary. (2007). DBpedia: A Nucleus for a Web of Open Data. Lecture Notes in Computer Science. 6. 722–735.