World Views - A Digital Archive Infrastructure for the Georg Eckert Institute for International Textbook Research

Lena-Luise Stahn[1], Steffen Hennicke[1], Ernesto William De Luca[1]

[1]Georg Eckert Institute Leibniz Institute for International Textbook Research

{stahn,hennicke,deluca}@leibniz-gei.de

Abstract This paper outlines the aims of the newly established project World Views. The paper mainly shows the work still to be done, as the project has only recently started and is still in initial consolidation phase. It presents an overview of the planned infrastructure which will work as a digital archive in the textbook research field. Therefore a data middleware is to be implemented to enable integration and standardization of the various GEI data existing in diverse forms as well embed it in a broader semantic context, thus enabling “World Views”.

1 Introduction

Textbook research constitutes a rather diverse area of interest and research. By bringing together professionals engaged in textbook research and their manifold knowledge and expertise, the Georg Eckert Institute[1] (GEI) is the central research institution in this field. This role results in various new research projects which aim at promoting and using the textbook as a medium for research in the historical sciences generating large amounts of data which are especially characterized by their heterogeneity and furthermore, are bound to specific infrastructures tailored towards the different kinds of data requirements. It has become apparent that data curation has not been thoroughly considered in the projects’ workflows. Standardization and archiving strategies have mostly been neglected. Up to this day this has resulted in collections containing valuable and important data which, however, exist parallel in separate environments with mostly no interfaces or linking possibilities. This goes against the “Good Scientific Practice” [8] postulated by the German Research Society (DFG) as the projects were funded by public money and the projects’ data and results constitute valuable scientific knowledge worth of long-term curation and preservation so as to allow sustainable usage. Often no knowledge exists of the existing data even within the GEI thus resulting in double work. Even if data has been “discovered” by other researchers its reuse is often difficult or even impossible because of legacy or out-of-date data formats.

Additionally, this situation has a negative effect on information retrieval and reuse by external parties since the GEI data neither is interlinked nor is it enriched with information from external sources. Despite its rich diversity and variety in terms of the available research data the GEI infrastructure lacks semantic contextualization. The new project World Views nationally funded (BMBF) and started in February 2015, is an effort to engage with the aforementioned issues. The absence of joint data storage is considered the main cause for this situation, which led to the decision to focus on establishing a suitable infrastructure, where the data integration of each existing project is made possible and which additionally serves as a standardizing basis for future project environments, eventually also leading to the implementation of a long-term curation strategy. With one joint search index which will work on all project data simultaneously the improvement of information retrieval is intended.

[1] http://www.gei.de/home.html

Another question concerns the semantic enrichment of the (meta-) data, which already forms a common method in other research environments [10, 11, 12].

1.1 Related work

Guideline papers about research infrastructures and research data reuse in the Humanities were found in [11] and in several DINI papers, e. g., [4]. A more international view was found in [6]. In Germany mostly CLARIN-D [1], DARIAH-DE [10, 11, 12] are used to build large information infrastructures, which is why these system have to considered during the evaluation and decision process. In case the World Views infrastructure requires a generic framework apart from these facilities [5] and [3] will provide a basis for evaluating the repository software. The adaption and use of the DTABf is shown in [7], revealing some of the system's main advantages. The remainder of this paper is organized as follows: part 2 provides an overview of the status quo at the GEI and its various data collections giving two examples of the GEI data in order to provide a more precise picture of how diverse the ways are in which the data is handled (2.1 and 2.2). A summary of the drawbacks resulting from this infrastructure closes this section. Part 3 discusses the aims of World Views and concludes the paper with a summary of the steps which have been taken so far.

2 GEI data

An overview of the GEI systems and their technical specifics illustrates the big gap between current research and information infrastructure guidelines [13, p. 11] and the actual situation: Edumeres[2], the information and communication portal for international educational media research amongst other provides access to the GEI's working papers with manually edited metadata and papers held in PDF; edu.data holds information on textbook systems worldwide; edu.experts, a database for textbook research professionals. The structure is implemented in Typo3 where every module has its own search and browse ui, partly with its own website as well (edu.data with Typo3 backend, edu.experts planned as Semantic Media Wiki). The Curricula Workstation provides central access to German and international curricula and also aims to create an archive of curricula. As they are mostly printed the curricula need to be scanned and stored in a DSpace repository, whereas the metadata is manually exported from the library OPAC. Parallel to the OPAC the VuFind-based TextBookCat provides a search entry point for the text book collection, with additional facettes and its own Solr index. And also infrastructures and web representations resulting from scientific projects[3] form a big data pool, e.g. „Nuances" providing teaching materials in various multimedia-based forms, or „Children and their world" trying topic modeling and again own Solr index. As such project proposals require a suitable web-based presentation, every time a new one is implemented and accompanied by its own system which seldom corresponds to the existing infrastructures, a behavior not to be expected to change in the future. With the end of the project's financial support these systems cannot be maintained appropriately and form the institute's legacy data, e.g. „DeuFraMat".

2.1 Data example: GEI-Digital

The GEI hosts one of the biggest research libraries in the field of textbook research. One of the goals is to make this library fit for the future by giving the library a so-called hybrid profile

[2] http://www.edumeres.net/nc/en/information/home.html

[3] http://www.gei.de/en/projects/current-projects.html

through digitizing its content. For this purpose, the project GEI-Digital has been initiated, in which the conversion of historic German speaking holdings into a machine readable format is being undertaken. An adequate research corpus has been created which can be used in diverse research areas.

The presentation platform provides digital images (generated through external providers) and automatically generated full text recognition file (via OCR) of every book page, written mostly in Gothic type. With this a basis for Digital Humanities tools is facilitated: a couple of projects, as e.g. “Children and their world”, have started to use methods such as topic modeling on this corpus. In June 2015 the database contained ca. 3,500 digitized and indexed textbooks with a time-span from 1648 to 1918 (ca. 900.000 digitized single documents). The used metadata format is METS for structural data and MODS for bibliographic descriptions, accessible through the GEI’s OAI-PMH interface. Data Integration takes place, amongst others, in Europeana[4] and the Deutsche Digitale Bibliothek[5] (DDB).

The corpus can be searched on metadata level as well as on full text level. Also a facetted search is provided, using the collection division which is based on metadata especially created for GEI-Digital, describing the type (atlases or storybooks) or subject of the textbook (geography, history), plus time frame in which it was used. Browsing options include the common bibliographic data. For the visualization of the digitized images the intranda viewer[6] is used. For each digitized document additional (meta-)data like ToC, thumbnail gallery, bibliographic data (partly also in English) and full text are provided and can be downloaded as METS/XML, MARCXML and DC via the OAI interface, Europeana Semantic Elements (ESE), OPAC/PICA, and PDF. Figure 1 shows an example screenshot. The Open Source Software Goobi[7] is used for digitization which provides an adequate environment for workflow handling and metadata editing. The metadata profile is generic; the Goobi interface has been customized and is filled manually except for the bibliographic data, which is harvested through the OPAC interface, thus also using the Gemeinsame Normdatei[8] (GND) data and Handle[9] service provided by GBV Common Library Network[10]. The same applies for the Solr index which is built at the GBV and adapted for GEI use. For backup the GEI cooperates with the Gauß-IT-Zentrum at the TU Braunschweig, where it holds server and storage capacities. Within this contract also the long-term preservation of the GEI-Digital data (i.e. digitized images, derivatives, metadata) is ensured.

2.2 Data example: project EurViews[11]

Through a comprehensive selection a collection of texts, maps and images from 20th and 21st century textbooks is established with the intention to present which notions of Europe and Europeans are conveyed through national textbooks. Historical and contemporary textbook sources from all European and many non-European countries are being incorporated, furnished with commentaries and contextualized information like histories of education, both of which

[4] http://www.europeana.eu/portal/

[5] https://www.deutsche-digitale-bibliothek.de/

[6] https://www.intranda.com/digiverso/intranda-viewer/intranda-viewer-overview/

[7] http://www.goobi.org/en/

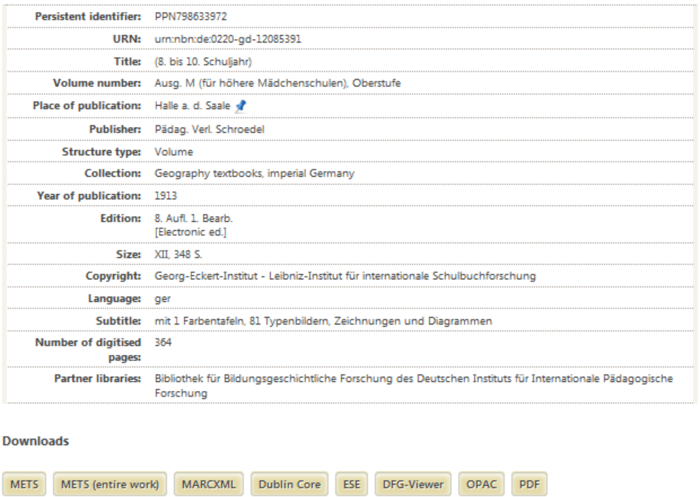
[8] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html

[9] http://www.handle.net/

[10] https://www.gbv.de/?set_language=en

[11] http://www.eurviews.eu/nc/start.html

are written by external researchers. Translations in German, English and sometimes also French or Spanish are provided by the EurViews members.

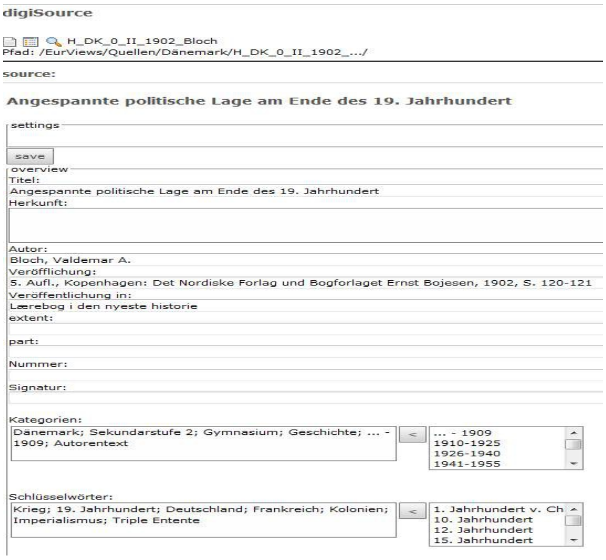


Persistent identifier:	PPN798633972
URN:	urn:nbn:de:0220-gd-12085391
Title:	(8. bis 10. Schuljahr)
Volume number:	Ausg. M (für höhere Mädchenschulen), Oberstufe
Place of publication:	Halle a. d. Saale
Publisher:	Pädag. Verl. Schroedel
Structure type:	Volume
Collection:	Geography textbooks, imperial Germany
Year of publication:	1913
Edition:	8. Aufl. 1. Bearb. [Electronic ed.]
Size:	XII, 348 S.
Copyright:	Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung
Language:	ger
Subtitle:	mit 1 Farbentafeln, 81 Typenbildern, Zeichnungen und Diagrammen
Number of digitised pages:	364
Partner libraries:	Bibliothek für Bildungsgeschichtliche Forschung des Deutschen Instituts für Internationale Pädagogische Forschung

Fig. 1. Bibliographic data view in Intranda Viewer

Although the workflow is similar, EurViews uses Typo3 as its digitization and metadata editing backend which is completely different from the one used in GEI-Digital. The homemade DigiSource extension supports handling the digitized image, storing the (meta-)data in a MySQL database, even though they are mostly already stored in GEI-Digital. The backup works through the TU Braunschweig, however no Solr index is used. The textbook sources can be searched on metadata level and on full text level. Additionally the sources are indexed with predefined search terms for time spans, categories, and keywords. A short summary completes each item's description. The collection is also accessible through a facetted search which clusters time periods, countries, and source types (i.e. structure type of the document). Figure 2 shows a screenshot of the Typo3 backend.

Both projects show the main problem to be handled: the data storage is done in a separate way, depending partly on old fashioned software. Especially the EurViews backend now turns out to be outdated resulting in unexpected high maintenance. The separate search indices prevent a comprehensive information retrieval. Instead of being of any help for the information seeking textbook researcher the separate search entry points constitute an obstacle which most people are not willing or able to overcome. Therefore, usability improvements as well as better information retrieval facilities are needed. Furthermore, there exists no connection to other GEI data: the sources' full texts and annotations as well as commentaries could serve as a knowledge base, easily enriched using other GEI data. However, lack of appropriate interfaces prevents the use by other GEI projects. Also the custom-built Typo3 extension prohibits data reuse. The metadata schema is generic and manually filled, despite the possibility of using bibliographic data from the library OPAC.



digiSource

H_DK_0_II_1902_Bloch
Pfad: /EurViews/Quellen/Dänemark/H_DK_0_II_1902_.../

source:

Angespannte politische Lage am Ende des 19. Jahrhundert

settings

overview
Titel:
Angespannte politische Lage am Ende des 19. Jahrhundert
Herkunft:

Autor:
Bloch, Valdemar A.
Veröfflichung:
5. Aufl., Kopenhagen: Det Nordiske Forlag und Bogforlaget Ernst Bojesen, 1902, S. 120-121
Veröffentlichung in:
Lærebog i den nyeste historie
extent:

part:

Nummer:

Signatur:

Kategorien:
Dänemark; Sekundarstufe 2; Gymnasium; Geschichte; ... - 1909; Autorentext
... - 1909
1910-1925
1926-1940
1941-1955

Schlüsselwörter:
Krieg; 19. Jahrhundert; Deutschland; Frankreich; Kolonien; Imperialismus; Triple Entente
1. Jahrhundert v. Ch
10. Jahrhundert
12. Jahrhundert
15. Jahrhundert

Fig. 2. EurViews metadata editing backend in Typo3

Analogous ways into the context of LOD/Semantic Web are absent: where GEI-Digital barely uses the controlled vocabularies provided by the GND, the EurViews data lack semantic contextualization completely.

3 Aims of World Views and first steps

The aforementioned projects are just two examples of the disconnected character of the collections, which in spite of its frequent use, reproduces this data and stores it several more times. Additionally there is the problem of the format variety: difficult to maintain, to curate and preserve, and almost every new project using a new data format. This leads to the emerging problem of data inaccessibility and thereby data loss. The aim of World Views is therefore to consolidate the various data sources. Based on a three-tier architecture model the project's focus is on implementing a central middleware, which will serve as the logic tier, where the metadata integration and standardization will be executed. The distributed and separated GEI data (this forms the data tier in the architecture) in its various formats will be drawn together, migrated into more standardized formats using metadata crosswalks and semantically enriched using internal and external links. Eventually the construction of one joint search Index (possibly based on Lucene/Solr) is planned, providing a comprehensive search through the main retrieval platform edumeres, additionally to the presentation on each project's platform, thus forming the infrastructure's presentation layer. Since the project has only just started no technical decisions have been made so far. To make learned decisions, comprehensive knowledge about the GEI's infrastructure and projects needed to be gained first. The project started with evaluation of main technologies available at the moment (open source as a requirement). Here mainly the product Fedora as the leading solution has been tested. But also DSpace and several other software solutions, e.g. infrastructures in the context of CLARIN-D and DARIAH-DE, are taken into account. The evaluation process contains creating a catalogue of requirements, testing the software on a virtual machine on both requirements and data to be used and comprehensive documentation. Eventually a decision is planned in the coming months. The decision on the bibliographic metadata does not seem to be the problem (main formats like DC are in the focus, as well as METS/MODS, as being already used) and crosswalks can be easily implemented, the main work lies in deciding on the extent of annotation. Here the infrastructure's intended character of persistence and sustainable usage has to be given consideration as it shall provide interfaces also for future projects, whose focus and functionality cannot be determined yet. For this project's part cooperation with professionals in the textbook research field is essential, since they form the user community. Their requirements and possible future ways of use will be surveyed through workshops and evaluations of other Digital Humanities projects. Thereby World Views is intended to function as a platform, which prepares the data for further use in the DH context. TEI [2, 3] as the most promising and prevalent format is the main focus of analysis. It was mainly chosen for its applicability on annotated texts produced in the humanities. Also its large community is considered as a benefit. Questions like annotation functionalities and how to exploit them in the most adequate way for the resource textbook have to be answered, forming a main part of the project's scientific work, since no metadata formats especially focusing on textbooks seem to be publicly available. Standardization plays also an obligatory part in the repository certification process, for example to achieve the DINI certificate [9], as is planned for World Views in the long run. The GEI data still resides in a mostly isolated position which stands in the way of representing the different "views", as the project title claims them. Therefore a vital point of World Views is the data contextualization. To get it enriched also provides its embedding in a semantic context which comprises interlinking the GEI projects further with controlled vocabularies up to Semantic Web applications, and thus providing comprehensive information retrieval possibilities as well as enriched corpora adequate for future DH research questions.

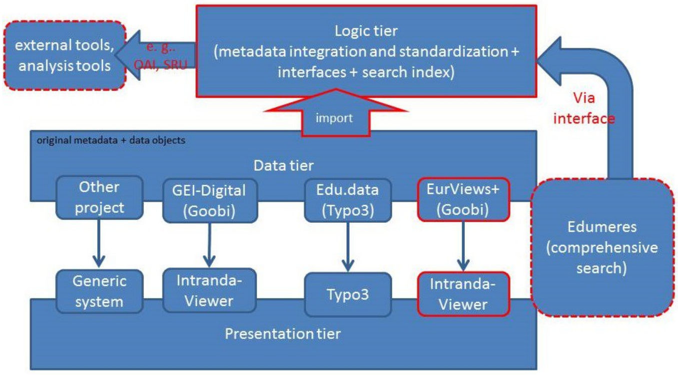

external tools, analysis tools
e. g., OAI, SRU
Logic tier (metadata integration and standardization + interfaces + search index)
Via interface
import
original metadata + data objects
Data tier
Other project
GEI-Digital (Goobi)
Edu.data (Typo3)
EurViews+ (Goobi)
Edumeres (comprehensive search)
Generic system
Intranda-Viewer
Typo3
Intranda-Viewer
Presentation tier

Fig. 3. World Views schematic representation

References

[1] Kommission Selbstkontrolle in der Wissenschaft Deutsche Forschungsgemeinschaft: Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft"; Denkschrift. Wiley-VCH (1998)

[2] DTA TEI Basisformat (2007-2015), http://www.deutschestextarchiv.de/doku/basisformat_en

[3] Dobratz, S. Open-Source-Software zur Realisierung von Institutionellen Repositories– Überblick. Humboldt-Universität zu Berlin, Zentraleinrichtung Universitätsbibliothek, Berlin (2007)

[4] Deutsche Initiative für Netzwerkinformation e.V.: Positionspapier Forschungsdaten. Arbeitsgruppe „Elektronisches Publizieren". (2009) http://edoc.hu-berlin.de/series/dini-schriften/2009-10/PDF/10.pdf

[5] Bagdanov, A., Katz, S., Nicolai, C., & Subirats, I.: Fedora Commons 3.0 versus DSpace 1.5: Selecting an enterprise-grade repository system for FAO of the United Nations. (2009)

[6] Battino Viterbo, P., Gourley, D.: Digital humanities and digital repositories: sustainable tech-nology for sustainable communications. In: Proceedings of the 28th ACM International Conference on Design of Communication (SIGDOC '10), pp.109-114. ACM, New York, NY (2010), USA, doi:10.1145/1878450.1878469

[7] Haaf, S., Geyken, A., Wiegand, F.: The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources (2012), doi: 10.4000/jtei.1114

[8] Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, Berlin (2012), http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf

[9] DINI: DINI-Zertifikat für Open-Access-Repositorien und-Publikationsdienste 2013 (2014), http://edoc.hu-berlin.de/series/dini-schriften/2013-3/PDF/3.pdf

[10] Beer, N., Herold, K., Kolbmann, W., Kollatz, Th., Romanello, M., Rose, S., Walkowski, N.-O.: Interdisciplinary Interoperability. DARIAH-DE Working Papers Nr. 3. DARIAH-DE, Göttingen (2014), urn:nbn:de:gbv:7-dariah-2014-1-0

[11] Fiedler, N., Werthmann, A., Stührenberg, M., Schonefeld, O., Bingel, J., & Witt, A.: Forschungsinfrastrukturen in außeruniversitären Forschungseinrichtungen: Forschungsbericht. (2014), http://dok.ids-mannheim.de/xmlui/bitstream/handle/10932/00-0230-5FEB-262D-CA01-8/Forschungsinfrastrukturen.pdf?sequence=4

[12] Puhl,J., Andorfer, P., Höckendorff, M., Schmunk, St., Stiller, J., Thoden, K.: Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften. DARIAH-DE Working Papers Nr. 11. DARIAH-DE, Göttingen (2015), urn:nbn:de:gbv:7-dariah-2015-4-4

[13] Research Infrastructures in the Leibniz Association (2015), http://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Presse/Publikationen/Leibniz_Infrastrukturen_2-2015_web.pdf