Modeling scholarly publications for sustainable workflows

Klaus Thoden

The project is currently funded by the German Federal Ministry of Education and Research, Grant Number 16OA061.

Introduction

1 The publication platform of the Edition Open Access[1] has been publishing scholarly monographs in a hybrid way since 2010, which means that the publications are disseminated in multiple digital formats (PDF, EPUB and HTML), as well as through a print-on-demand service as printed books. Currently, a new version of the platform is designed and developed, as available technology has moved on significantly since the first version. Also, the original infrastructure did not adhere much to existing standards, but rather relied on home-grown and proprietary formats. There are many reasons, though, why existing standards should be used to a greater extent. Firstly, the format of the publications should be one that can be re-used in other project contents. Secondly, it should be possible for other projects to set up an instance of this publication platform and feed it with their own material. And thirdly, from the standpoint of interoperability as well as file-preservation strategies, existing and open standards should be chosen.

2 Since the new platform will eventually also host the 30 existing publications, a method needs to be found to transfer the corpus of texts from the original, proprietary format to a sustainable and standardized data format. This transfer is not without difficulty, though, since the development of the workflow was concurrent to the development of the whole platform. Thus, the source files of the early books were originally in a format that is not compatible anymore with the current document conversion workflow. Also, in the days before version control systems became mainstream some source files have been lost, or it is unclear if the available files reflect the final version of a publication.

3 Consequently, the aim of this study is to show how a publication project can benefit from a clearly defined document model, so that all stages of the workflow can be monitored for

consistency. The paper is organised at follows. First, the significance of the study is made clear, followed by a description of the methodology and a comparison between the available standards. Finally, the problems pertaining to the specific project are addressed and solutions are shown, where possible.

Key objectives of the study and significance

4 From the original design, the Edition Open Access planned to release its content in multiple formats: PDF, EPUB and HTML as well as a print-on-demand paper version. Although the majority of scholarly publications is still published as PDF, the format has its disadvantages, especially concerning the ways it can be re-purposed for research, other than being downloaded and annotated, since it is merely the electronic version of work meant to make hard copies from. We claim that steps need to be taken to move forward and open up the space for a variety of formats, each of which has their specific purpose. One example makes this quite apparent: viewing a PDF document on a smart phone screen does not give the user a good reading experience, while an EPUB file or a webpage with responsive design will render the contents adapted to such a small screen.

5 Electronic publishing in Open Access creates new ways in displaying publications and enables linking of a publication with its sources or with supplemental material. It lowers the boundaries between different knowledge bases and removes disruptions like having to move from a digital medium to an analogue one or hitting upon a paywall. For creating an unbounded web of knowledge, also electronic publications need to be integrated seamlessly into the world wide web. And the format for this is presently a representation in HTML.

6 The ways in which a scholarly publication can be used are immensely increased if they can be re-purposed not only as being read, but also if the structured contents can be exploited, for example through creating spin-off documents by extracting only the Latin text of a bilingual source edition, by enriching the text with information gained from entity extraction or by performing citation network analysis. Besides its prosaic nature in documenting research, the text can also be seen as a representation of research data. And being available in a structured format, it will be highly useful once the publication itself becomes the object of research. A shift of thinking in data instead of paper has to be made.

Methodology

7 In the realm of single-source publishing it is crucial that there is one format from which the resulting output formats can be easily derived. This is quite a complex arrangement that depends on a strict and closed set of available phenomena that are expected in the documents, e.g. the kind of markup, floating elements or formulæ. Only if it is foreseeable what elements are allowed and expected can conversion tools be adapted or developed that create the desired output.

8 The best means to define this kind of structure is an abstract model of a scholarly publication. In the case of Edition Open Access, most publications deal with the topic of history of science. Although in and of itself a humanities subject, the topics that are dealt with can easily be situated in the sciences, so mathematical or chemical formulæ can be

expected as well as facsimile pages and critical text editions. Thus, in the document model that is to be created, these features have to be taken into account. Furthermore, the document model lives independent of the contents of a publication, and serves as the blueprint for a monograph: it is used to document the available components and their dependencies (cf. Maler/ El Andaloussi 1996).

9 For the current project, the model is realized following the P5 Guidelines of the Text Encoding Initiative (TEI), a quite universal markup format that can be used in a lot of ways (Burnard/ Rahtz 2004; Odebrecht 2015). Its format of documentation, ODD ("One Document Does It All"), is designed to contain the user manual for the editorial staff as well as code for validating XML files. Depending on how the ODD is compiled, it either creates the editorial manual as HTML or PDF files or a rule-based schema (e.g. W3C Schema, Relax NG) against which an instance of the model, that is, a scholarly publication can be checked and validated. This method of intertwining code and documentation of code can be traced back to the *Literate Programming* approach as described in Knuth (1992). This approach has been put to use e.g. in the documentation of LaTeX packages (Pakin, 2015) and more has recently gained momentum in the context of reproducible research, as described in Schulte/ Davison (2011) or Thoden (2018).

State of the art in XML-first publishing

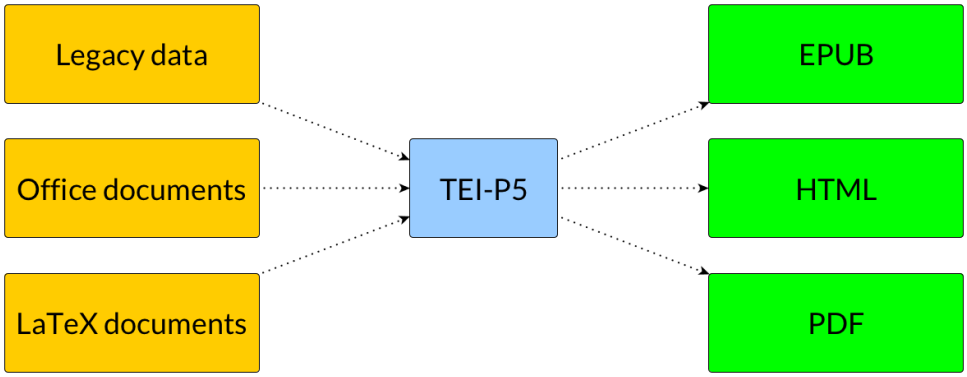
10 The choice using TEI-P5 ODD as the modeling language is in line with the XML format that is going to be used for encoding the scholarly publications. In the field of XML-first publishing, two standards are currently prevalent: NLM-JATS (NISO, 2019) (and its customization for books, BITS) and the Guidelines of the Text Encoding Initiative (The TEI Consortium, 2019). While JATS/BITS originates from the sciences and has been standardized by NISO, The TEI Guidelines were developed in the context of creating digital editions in the humanities.

11 Both approaches have in common that they contain not exactly one schema. JATS, for example offers three different tag sets, while with TEI, each project is encouraged to create its own customization of the Guidelines.[2] Considering these variations, it becomes all the more apparent, that, whichever system a particular project intends to use, the documentation of its methods and the underlying data structure are key elements in creating a sustainable environment in which publications are created. Following a set of open standards and documenting its use in a project also paves the way for creating or re-using conversion tools to move from one format to another.

12 Holmes/Romary (2011) argue for the possibility of adapting TEI as a format for scholarly publications, and it has been adopted by endeavours such as DHQ (Flanders et al. 2015) and Lodel (Terrier 2018). In the case of the Edition Open Access, the decision to use TEI was made, partly because one series of publications comprises editions of rare books. Since the original files of digital editions are already present as TEI-P5 documents, the complete or partial inclusion of the source material into a publication is made easy.

Hands-on experience

Workflow schema



(illustration made by the author)

13 Besides the multiple output formats offered by Edtion Open Access there are also several input formats that need to be taken care of: two standard workflows (office documents and LaTeX), as well as the recovery of the back-catalogue.

14 On the new platform, TEI-P5 documents will be used as the central data format. All different input formats will be converted to TEI-P5, and the structure of the documents is checked against the document model described above. The documentation in the ODD file also serves as a manual for the editorial staff.

15 In accordance with the single-source publishing approach, the TEI-P5 file is converted to the desired output formats. The figure above shows a schematic overview of the different ways of input and output documents. The following sections highlight the practical problems in the three parts of the workflow and also describe how these problems are handled.

Creating an abstract document model

16 Creating a document model is an iterative process. In the case of the current project, the existing publications were used as the first piece of input. Following the available documentation,[3] an example TEI document was created to see how each element could be encoded according to the TEI Guidelines. Making sure that all existing phenomena are covered is important for the backwards compatibility of the model. Additionally, more features were added. For example, epigraphs had been used in publications already,[4] although they never had been codified. Similarly, abstracts on chapter level were handled rather unofficially. Especially in the latter case, by marking up an abstract appropriately makes it possible to address this piece of the document and to re-use it in different contexts, e.g. the table of contents.

17 Thus, by re-iterating the encoding and checking against the existing publications, a stable version of an example document was created. This document was then used to create the document model in ODD, which also is not without complications.

18 The ODD format employs the concept of keeping documentation for humans and rules for computers synchronized in one file. One issue that has to be faced is the general structure of this format. In the ODD, the SCHEMASPEC element is used to define the document model, and elements to be used are included or excluded in the respective MODULEREF elements. Each of the elements can be further described and defined in the ELEMENTSPEC elements. But this only allows a general definition of the elements, without taking care of the context they are used in. What if an element is used in two different contexts and hence requires different restrictions? The solution here is to use the schema language Schematron (ISO/IEC 2016), whose rules can be integrated into ODD files and can be used to describe the various contextual behaviours of elements.

19 The overall question is whether or not a document model should be that strict anyway. After all, besides the need to have complete control over the workflow and addressing the various output formats, the model as well as the TEI files are supposed to be interoperable. One solution could be to employ a multi-level validation system. A very strict system, possibly using a customized Relax NG schema, could be used for a strict validation when editing the content. A slightly broader schema could be used for data interchange, documented in the ODD. As long as both schemata are a subset of the TEI Guidelines, this would be feasible, but care should be taken since two separate schemata would have to be maintained.

Conversion of legacy documents

20 As already mentioned, the establishment of the new web platform means that the existing publications need to be ingested into the new system, which will work directly with TEI documents, as opposed to the initial way which consisted of inserting the elements of an XML file into an SQL database structure. Since the format of the available source files is inconsistent, the decision was made to reconstruct these publications from the contents of the SQL database. Since the database schema of the publication platform has not seen any significant changes in its nine years of existence, the text of the publications has been preserved there and can be used as the basis for a TEI version. In the case of more recent publications, the situation is a lot clearer, since mechanisms like version control have been employed and all relevant files have been archived properly.

21 The database export is not without complications. Since the original files were originally written in LaTeX, bibliographical information was kept in a separate database and referenced in the source files via a shortcut. For the online version, the references were resolved to a properly formatted citation, but no hint regarding the shortcut was retained. During export, the connection between the formatted citations and their counterparts in the bibliographical database needs to be re-established heuristically and interactively.

22 Another feature of the publication platform is also that each paragraph can be addressed through a URL,[5] so that it can be used as a citation marker. During the process of exporting, refining the TEI document and subsequent uploading into the new platform, it has to be made sure that the old URL handles will still point to the correct paragraph.

Producing output formats

23 The construction of the XML based output formats (EPUB, HTML version) is a rather straightforward operation and is handled mainly by XSLT and XQuery scripts. This is of advantage, since an HTML view of a publication can be generated at any time and can be given to people for reviewing.

24 Producing the PDF output has some complications. When committing the contents to a printed page, layout considerations need to be observed more closely, for example when placing floating objects like figures or tables on a page. Moreover, there are a number of ways to get from an XML document to a PDF version, the most obvious being XSL-FO. However, there are major differences in quality and power between a free version like Apache FOP[6] and commercial products. Secondly, desktop publishing programs like Adobe InDesign might be used, but there might be difficulties with typesetting formulæ. Ideally, the workflow should be flexible enough so that for each publication it should be possible to choose the most convenient tool. This could be dependent on the number of images and use of math.

25 In the context of the present project, the choice fell upon a TeX based typesetting system. One of its derivates, ConTeXt, is able to work directly with TEI-P5 data by means of a stylesheet that contains directives for the rendering of elements (Schmitz 2011). However, as the only output of the typesetting process is a PDF file, there is no way of having case by case decisions concerning the placement of floating objects or the control of line or page breaks. While these cases could be dealt with by using processing instructions in the TEI file it would introduce format specific instructions into the TEI document, which should be avoided. In the end, a solution was chosen where an XSL script converts the TEI-P5 data to common LaTeX files out of which the PDF is generated. This step guarantees full control over layout issues, but it needs to be taken care of that any edits to the textual content are made in the master TEI file.

Conclusion

26 This paper explores the way of how a running publication project deals with its legacy in firstly creating an abstract document model, that will deal with all the phenomena that are present in the books, and then to devise a system for salvaging the legacy data, to re-map it to the document model and to ultimately develop a system that uses a standardized and well-documented basis for its past and future publications. Even though the source files of some books have become incompatible with the current system, the way they are stored in the database for serving the online version has remained stable.

27 This study shows above all that the creation of an underlying data model is very important and needs to be well-documented and well-maintained to serve its diverse purposes, amongst them providing encoding guidelines for the editorial staff, as well as serving as a rule-based schema for document validation. The TEI format is capable of being the pivot point for representing the master version of a publication and allowing the conversion into all sorts of formats. This also goes with the fact that output formats and viewing applications will change much more rapidly than the underlying data needs to be changed. And in this case it is crucial that there is one definite and stable format that can be transformed into the desired output.

28 Ultimately, the choice of standard is dependent on a lot of factors, ranging from the standards favoured in a scholarly discipline or world region to the integration of available material, that already exists in one particular format. In that sense, the concept of bibliodiversity also mirrors in this work, as one source, the documentation of the work of scholars is released into the eco-system of publishing and can interact with a wide range of actors, both man and machine.

BIBLIOGRAPHY

References

Burnard, Lou, and Sebastian Rahtz. 2004. 'RelaxNG with Son of ODD'. Presented at the Extreme Markup Languages 2004,Montreal, QC, Canada, August 2.

Flanders, Julia, Wendell Piez, John Walsh, and Melissa Terras. 2015. 'Challenges of an XML-Based Open-Access Journal: Digital Humanities Quarterly'. Presented at the (Proceedings) Digital Humanities 2015. http://dh2015.org/abstracts/xml/FLANDERS_Julia_Challenges_of_an_XML_based_Open_Ac/FLANDERS_Julia_Challenges_of_an_XML_based_Open_Access_J.html.

Holmes, Martin, and Laurent Romary. 2011. 'Encoding Models for Scholarly Literature'. In Handbook of Research on Green ICT: Technology, Business and Social Perspectives, edited by Ioannis Iglezakis, Tatiana-Eleni Synodinou, and Sarantos Kapidakis. IGI Global. https://doi.org/10.4018/978-1-61692-834-6.ch005.

ISO/IEC. 2016. Information Technology — Document Schema Definition Languages (DSDL) — Part 3: Rule-Based Validation, Schematron, International Standard ISO/IEC 19757-3. Geneva, Switzerland: ISO. https://www.iso.org/standard/55982.html.

Knuth, Donald Ervin. 1992. *Literate Programming. CSLI Lecture Notes, no. 27*. Stanford, Calif.: Center for the Study of Language and Information.

NISO. 2019. *JATS: Journal Article Tag Suite, Version 1.2 (ANSI/NISO Z39.96-2019)*. Baltimore, Maryland, U.S.A: National Information Standards Organization.

Maler, Eve, and Jeanne El Andaloussi. 1996. *Developing SGML DTDs: From Text to Model to Markup*. Upper Saddle River, N.J: Prentice Hall PTR.

Odebrecht, Carolin. 2015. 'Interdisziplinäre Nutzung von Forschungsdaten Mithilfe Einer Technisch-Abstrakten Modellierung'. Presented at the Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum, Graz.

Pakin, Scott. 2015. 'How to Package Your LaTeX Package'. https://www.ctan.org/pkg/dtxtut.

Schmitz, Thomas. 2011. 'Kritische Editionen mit TEI Xml und ConTeXt setzen'. presented at the Frühjahrstagung von DANTE e.V., Bremen. http://www.dante.de/events/Archiv/dante2011/programm/vortraege/folien-ts.pdf.

Schulte, E., and D. Davison. 2011. 'Active Documents with Org-Mode'. *Computing in Science Engineering* 13 (3): 66–73. https://doi.org/10.1109/MCSE.2011.41.

Terrier, Caroline. 2018. Creating a TEI Document in Lodel 1.0. OpenEdition TEI Schema. (version 3256e57). OpenEdition. https://github.com/OpenEdition/tei.openedition.

Thoden, Klaus. 2018. 'Interaktives Publizieren im Open Access'. presented at the Open-Access-Tage, Graz, September 25. https://doi.org/10.5281/zenodo.1441084.

The TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

NOTES

1. http://edition-open-access.de (checked March 22, 2019)
2. https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/(checked March 22, 2019)
3. http://edition-open-access.de/media/support/files/EOAReference.pdf (checked March 22, 2019)
4. e. g. http://mprl-series.mpg.de/media/studies/8/8/Studies8chap7.pdf (checked March 22, 2019)
5. e.g. http://www.edition-open-sources.org/sources/10/7/index.html#2 (checked March 22, 2019)
6. https://xmlgraphics.apache.org/fop/ (checked March 22, 2019)

ABSTRACT

This study deals with the strategy of converting the workflow and document basis from a proprietary format to a fully standards-compliant system in the context of a publishing platform, that offers multiple output formats of monographs in the arts and humanities. It stresses the importance of creating an abstract document model as the basis for this single-source publishing approach and how a model offers guidance on each step of the way in book production.

INDEX

Keywords: Literate programming, legacy data, single source publishing, XML, workflow

AUTHOR

KLAUS THODEN
Max Planck Institute for the History of Science
kthoden@mpiwg-berlin.mpg.de