Journal of the Text Encoding Initiative

Issue 13 | May 2020 - November 2022
Selected Papers from the 2018 TEI Conference

Digital Texts in Practice

Christian Wittern

Electronic version

URL: <https://journals.openedition.org/jtei/3187>
DOI: 10.4000/jtei.3187
ISSN: 2162-5603

Publisher

TEI Consortium

Electronic reference

Christian Wittern, "Digital Texts in Practice", *Journal of the Text Encoding Initiative* [Online], Issue 13 | May 2020 - November 2022, Online since 19 October 2020, connection on 25 November 2022. URL: <http://journals.openedition.org/jtei/3187> ; DOI: <https://doi.org/10.4000/jtei.3187>

For this publication a Creative Commons Attribution 4.0 International license has been granted by the author(s) who retain full copyright.

Digital Texts in Practice

Christian Wittern

ABSTRACT

As a student of intellectual, religious, and cultural developments in areas of the Chinese cultural sphere, my initial motivation for engaging with digital texts thirty years ago was to open up the new possibilities that the digital medium offered to researchers, without losing any of the affordances of a traditional printed edition. This requirement includes use of texts for reading, translating, annotating, quoting, and publishing, thus integrating with the whole of the scholarly workflow.

At that time theories of electronic texts started to appear and the Text Encoding Initiative had already begun to create a common text model and interchange specification, based mainly on European languages. For East Asian texts, things were much more complicated because of different and quickly evolving character encoding standards, different textual traditions and approaches to text editing, as well as different institutional embedding.

In this paper, I will look back at these developments, first to recount some of the history, albeit from a strictly personal perspective, but also to take stock of the situation and consider where we are now, how we got there, and what remains to be done to realize the dream of the universal digital text, easily shared and annotated, but still tractable, verifiable, and authoritative.

INDEX

Keywords: text encoding, character encoding, Chinese texts, SGML, XML, TEI

AUTHOR'S NOTES

This paper was originally delivered as a keynote presentation at the TEI meeting in Tokyo in September 2018. I would like to thank the audience for the interesting discussion and feedback. In reworking the presentation into this paper, I have tried to take this into account, while preserving traces of the oral delivery.

1. Introduction

- 1 In this paper, I will look back at different stages of my work with digital texts. I do this not so much because I consider what I did particularly important, but rather because I think it represents some important milestones in the development of digital texts, the theoretical basis for them, and ways of working with them. And of course I do it because I know a bit more about my own work than about other things, even if these other things might be more important.
- 2 I am a Sinologist by training; my interest lies mostly in Buddhist and Daoist texts from China of the eighth to twelfth centuries, but also includes intellectual history and poetry of many periods. I stumbled into what is now called digital humanities about the time I was starting to write my master's thesis, because I assumed that the computer would be a shortcut to working with the huge amount of material. Rather than reading all the stuff myself, I thought it would be easy to let the computer do the hard work and then reap the rewards. Boy, was I wrong! Instead, for many years I had almost no time to read books because there was so much missing on the computer side that I could only try to make up for that by filling in the blanks (or at least participating in the filling in) myself.

- 3 My real vision was to be able to work with digital texts in the same way as with printed texts, but on top of that *also* being able to search, analyze, annotate, excerpt, and reorganize the content. What I mean by “in the same way” specifically includes things such as carrying the texts around with me (even without access to a network, which at that time mostly connected mainframe computers over dial-up lines, mainly for military use) and being able to scribble in the margins to keep a record of what I had already read and what I thought of it. And of course the digital text would have to be as “easy to use” as a book.

2. First Steps: Encoding and “Missing Characters”

- 4 My work with digital texts can be traced back to a rainy afternoon in October 1989, when I opened in my Hamburg home the laptop computer I had acquired in Hong Kong on my way back from a two-year stay at Hangzhou University, People’s Republic of China, and finally managed to switch the display to Chinese characters. At that time the computer was running the English version of MS-DOS 3.3, but clever programmers had found a way to overlay the screen with Chinese characters through a memory resident overlay (see [figure 2](#)).

Figure 1. The ETen Chinese System along with the hardware extension card for desktop computers.

- 5 However, right here I encountered my first difficulty. The laptop I had bought did not have the standard VGA resolution of 640 x 480, but had the slightly different resolution of the Hercules graphics card, which should be 720 x 350. However, the display of this laptop had only an emulation of the Hercules graphics mode, which displayed 720 x 480 dots on the screen. The software I was using, ETen Chinese System (倚天中文系統; see figure 1), was designed for either the standard VGA or the Hercules mode, but to start it for this specific hardware, it needed a special command line switch that was not documented in the manual.

Figure 2. Startup screen of the Chinese system.

- 6 In the previous paragraph, you may have noticed many highly specialized terms that even then were hardly known among technically minded people, let alone aspiring Sinologists. Since I had absolutely no knowledge of the workings of computers or how to operate them, it took me six weeks to get far enough in the manual, which was written in Chinese, to figure out how to start the Chinese mode. I had to find the batch file which contained the appropriate command-line switch before I could type Chinese characters. I then immediately started to type in the poems I was working with for my master's thesis in order to create a critical edition, analyze them, and make a concordance. These were about three hundred poems from manuscripts found in the Central Asian oasis city of Dunhuang at the beginning of the twentieth century, with some fragments from collections in Japan (see [figure 3](#) for an example of such a manuscript fragment). Fortunately for me, there was already a modern edition, but it was not very reliable, and many of the characters in the manuscripts were not in the modern standard form used to create the fonts in the Chinese system I was using.

Figure 3. A page from manuscript Stein 778, a fragment of Wang Fanzhi's poems.

圖二 伯三二一一王梵志詩集卷中

- 7 Luckily, since I had studied the manual so thoroughly, I knew that there was a little utility program that allowed users to create additional characters if the standard ones did not suffice. Apparently, this problem was common enough to already have a ready-made solution. I learned that a separate region in the codespace had been allocated for the definition of such character shapes and they could even be added to the input system, thus making them first-class citizens indistinguishable from the other characters.

2.1 Codes and Coding Systems

- 8 At this point, it might be useful to very briefly introduce character encoding. Computers were invented in England, but the development into the kind of machines we use today was done in the United States. At first every system had its own way of assigning the internal representation, that is, numbers to visible shapes (i.e., characters). Since this was not sufficient when computers started talking to each other, a standard for encoding was needed. This came to be called the American Standard Code for Information Interchange (ASCII), which contained only the most

urgently needed characters (in the version of 1962 only the uppercase letters of the English alphabet; the lowercase rows were added in 1963), since memory allocation was an important consideration. ASCII assigns 127 slots (see figure 4), most to visual characters, but also a significant number to control characters. These slots are represented internally as a fixed sequence of “on” and “off” states, usually represented in writing as 1 and 0. The numbers are thus represented as integers with base two. A binary number with 7 digits can encode up to 128 slots. This came to be called a “byte,” while every individual digit is a “bit.” Internally the bits are represented in memory in fixed spaces of length 8, but one of the bits was used for control purposes, so only 7 bits were available for encoding in ASCII. All coding systems that have been developed since then bear some kind of relationship to ASCII: most involve an extension of the codespace; some also a slight change in the assignment of non-alphabetical character shapes. The system I happened to start my computing life with was developed in Taiwan and thus used the Big5 encoding, a coding system for traditional Chinese, which enlarges the ASCII codespace by using two ASCII characters to represent one Chinese character. Big5 is an industry-defined character encoding system (I avoid the word standard here, since there are actually a number of competing, slightly different versions and there is no definite standard document¹), while most of the other character encoding systems used in East Asia are defined by the government bodies charged with defining standards. Big5 also allows text in plain ASCII to mingle with Chinese, making it much easier to use in practice than the first batch of official encodings, which assume a completely localized system with no room for plain ASCII. Since characters can occupy either one position (English) or two positions (Chinese), the processing becomes more complex and requires specialized software.

Figure 4. ASCII code table (1963).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	00 0000 NUL	01 0001 SOH	02 0010 STX	03 0011 ETX	04 0100 EOT	05 0101 ENQ	06 0110 ACK	07 0111 BEL	08 1000 BS	09 1001 HT	0A 1010 LF	0B 1011 VT	0C 1100 FF	0D 1101 CR	0E 1110 SO	0F 1111 SI
1	10 0010 DLE	11 0011 DC1	12 0100 DC2	13 0101 DC3	14 0110 DC4	15 0111 NAK	16 1000 SYN	17 1001 ETB	18 1010 CAN	19 1011 EM	1A 1100 SUB	1B 1101 ESC	1C 1110 FS	1D 1111 GS	1E 1110 RS	1F 1111 US
2	20 0010 SP	21 0011 !	22 0100 "	23 0101 #	24 0110 \$	25 0111 %	26 1000 &	27 1001 '	28 1010 (	29 1011)	2A 1100 *	2B 1101 +	2C 1110 ,	2D 1111 -	2E 1110 .	2F 1111 /
3	30 0010 0	31 0011 1	32 0100 2	33 0101 3	34 0110 4	35 0111 5	36 1000 6	37 1001 7	38 1010 8	39 1011 9	3A 1100 :	3B 1101 ;	3C 1110 <	3D 1111 =	3E 1110 >	3F 1111 ?
4	40 0100 @	41 0101 A	42 0110 B	43 0111 C	44 1000 D	45 1001 E	46 1010 F	47 1011 G	48 1100 H	49 1101 I	4A 1110 J	4B 1111 K	4C 1110 L	4D 1111 M	4E 1110 N	4F 1111 O
5	50 0100 P	51 0101 Q	52 0110 R	53 0111 S	54 1000 T	55 1001 U	56 1010 V	57 1011 W	58 1100 X	59 1101 Y	5A 1110 Z	5B 1111 [	5C 1110 \]	5D 1111 ^	5E 1110 _	5F 1111 `
6	60 0100 a	61 0101 b	62 0110 c	63 0111 d	64 1000 e	65 1001 f	66 1010 g	67 1011 h	68 1100 i	69 1101 j	6A 1110 k	6B 1111 l	6C 1110 m	6D 1111 n	6E 1110 o	6F 1111 p
7	70 0100 q	71 0101 r	72 0110 s	73 0111 t	74 1000 u	75 1001 v	76 1010 w	77 1011 x	78 1100 y	79 1101 z	7A 1110 {	7B 1111 	7C 1110 }	7D 1111 ~	7E 1110 DEL	7F 1111 DEL

2.2 The Missing Characters

- 9 Back to the poems for my master's thesis. I could type and display all the characters easily, and even print them with my dot-matrix printer. But later, when I started to use the quite nicer outline fonts that allow for much more detailed renderings of the characters, which is especially necessary for kanji, I discovered that these customized characters were missing, since there was no way to add them to an existing font. That was the first time I encountered a problem I would spend a lot of time with in later years: attempting to make these privately defined characters in some ways interchangeable across boundaries of character encodings, operating systems, users, and language environments.

3. ZenBase

- 10 After receiving my master's degree, I set out for graduate studies in Kyoto and soon found myself busy helping Urs App and his Zen Knowledgebase project at the International Research Institute for Zen Buddhism of Hanazono University. He was trying to create a text database of all the texts

that played a role in the development of Zen (or Chan) Buddhism and was struggling with problems similar to mine. More importantly, he was also putting out a newsletter, the *Electronic Bodhidharma*,² which tried to connect projects working in the field of Buddhist studies, make them compatible across language barriers, and also encourage collaboration and avoid duplication.

- 11 Urs was also a founding member of the Electronic Buddhist Text Initiative (EBTI), which was started by Lew Lancaster at UC Berkeley in 1993. Lew travels a lot and picks up on new trends early. One souvenir of a visit he made to our Kyoto Institute that year was a thick tome titled *The SGML Handbook* (Goldfarb and Rubinsky 1990), which landed on my desk and claimed to be important. I started to dig into it, but it proved to be very hard to read. Looking for easier access points and other people struggling to understand SGML, I found references to something called the “Text Encoding Initiative” and discussions of “P2,” which appeared to be a set of draft documents on an FTP server in Oslo, Norway. This was a real revelation and I immediately saw that TEI would be the future of digital texts: “These Guidelines ... are addressed to anyone who works with any text in electronic form. They provide a means of encoding those features of a text which need to be identified in some way in order to aid the processing of that text by computer programs”³ (Sperberg-McQueen and Burnard 1993). And this information about the documents could also be defined in a machine-readable and machine-exchangeable form! Eureka! Another result of this discovery was that we decided we needed an introduction to TEI and invited the editors to the next EBTI conference, which was being held at Haienji, a remote mountain monastery in Korea, in October 1994, to discuss a “Buddhist DTD” (as we then called it, without really understanding what a DTD was or why it was useful).
- 12 As you might expect, there were also many setbacks, especially since applying TEI to East Asian texts with their various encodings and different writing conventions was still quite a tall order, but the connection was made, and for the next twenty years, applying TEI to East Asian Texts formed a considerable part of my professional life. Urs understood the necessity of all this markup and SGML very well (and wrote enthusiastically about it in the newsletter: see App 1993; 1995a, b, c), but he was still reluctant to burden our users with it right away, since a toolchain for working with these SGML encoded texts was still largely missing and what software was available would not work with our Chinese texts. So the texts on the *ZenBase CD 1* (App, Wittern, and Fujimoto 1995), which was published in June 1995, were still plain text without TEI markup, but each of them had a TEI

header attached, because we realized that having the metadata of the texts in machine-readable form was itself a big win. I did manage to sneak in one text I had encoded in TEI P3 (see figure 5), just to demonstrate the usefulness of this new technology, but we needed to provide a program that would strip out the markup to make it usable in the same way as the other texts. Use of the CD depended upon a grep-like tool, which extracts matching lines from a text file, to locate a search term in the text.

Figure 5. TEI header of the first published Chinese text marked in TEI, the Wudeng Huiyuan (ZenBase CD 1, June 1995).

```

<!DOCTYPE TEI.2 PUBLIC "-//TEI//DTD TEI CWSmall 1.0//EN"
[
  <!Entity % ENTY SYSTEM "wdhygif.ent" >
    %ENTY;
  ]>
  <TEI.2>
  <TEIHEADER>
  <FILEDESC>
  <TITLESTMT>
  <TITLE>ELECTRONIC TEXT OF THE WUDENG HUIYUAN (JAP. GOTO EGEN)</TITLE>
  <PRINCIPAL>Responsible Editor: Urs APP</PRINCIPAL>
  <RESPSTMT><NAME>Christian WITTERN</NAME>
  <RESP>Conversion to TEI-conformant markup.</RESP></RESPSTMT>
  </TITLESTMT>
  <EDITIONSTMT><EDITION>BIG5 Code version 1.0</EDITION></EDITIONSTMT>
  <PUBLICATIONSTMT><PUBLISHER>International Research Institute for Zen Buddhism</PUBLISHER>
  <ADDRESS>
  <ADDRLINE>Hanazono University, Tsubo no uchi cho 8-1, Nishinokyo</ADDRLINE>
  <ADDRLINE>Nakakyo-ku, Kyoto, 604 JAPAN Fax (075) 811-9664</ADDRLINE>
  </ADDRESS>
  <AVAILABILITY><P>free</P></AVAILABILITY>
  <DATE>June 1, 1995</DATE>
  <DISTRIBUTOR>International Research Institute for Zen Buddhism</DISTRIBUTOR>
  </PUBLICATIONSTMT>
  <SOURCEDESC>
  <BIBL>PU J1: Wudeng Huiyuan, Zhonghua Shuju ed. 1984 ISBN 7-101-00536-5</BIBL></SOURCEDESC></FILEDESC>
  <ENCODINGDESC>
  <EDITORIALDECL>
  <P>The hierarchical division of the text has been represented by DIV1, DIV2 and DIV3 elements. DIV1 marks the chapter, DIV2
  (type="hokei") groups the disciples of a certain master, and DIV3 (type="zoku") is an individual entry.</P><P>The DIV3 elements are
  followed by a HEAD element, which gives the name and an ID code. The ID code has been taken from the ID-database developed at the IRIZ
  (canonical references of the form C-XX-YY-ZZ, where XX is the page, YY the row and ZZ the column of the entry in the lineage chart of the
  <TITLE>Zengaku Daijiten</TITLE>). Where this code did not apply, or could not be found, a unique ID of the form WXXX-YYY has been used,
  where XX is the chapter and YYY is the sequential number of the DIV3 in question. References to the master whose hagiography is reported
  are marked with the RS element where the KEY attribute references this same ID.</P><P>The Q tag has been used to mark direct speech, the
  QUOTE tag to mark quotations from other texts or citation of poems.</P><P>NAME tags have been used for proper names of persons and
  places. Occasionally dynastical dates are still marked as NAME, although they are supposed to be marked as DATERANGE.</P>
  <P>Textual variants are marked according to the notes supplied in the source edition without any further checking. The inline method with
  LEM and RDG elements within a APP element has been used. These elements bear a WIT attribute whenever indicated in the source, "Di"
  stands here for the base text (the one used by the editors of the Zhonghua Shuju edition), "ZZ" is the text from the Dainippon Zokuzokyo
  Vol 138, and "Q" is the Qing edition.</P></EDITORIALDECL>
  <TAGSDECL></TAGSDECL>
  <REFSDECL><P>Milestone tags are used to mark page- and line breaks in the source. PB is used for the page break, LB for the linebreak,
  with the ED attribute set to "ZHSJ" and the N attribut giving the number of page or line.</P></REFSDECL></ENCODINGDESC>
  </TEIHEADER>
  <PROTEIENES>
  B(DOS) ---WDHYB.SGM Top (1,0) (SGML -3 Wrap) Fri Sep 7 4:29PM 0.20
  Quit

```

- 13 Although a primitive method limited in many ways, this tool had made digital texts usable in practice for many scholars as far back as the early 1990s, and even today the regular expression tool grep is loved by colleagues of my generation in Chinese Buddhist Studies. And the TEI SGML-encoded version of the Wudeng Huiyuan (WDHY), despite the beauty of all these angled brackets, did not see much use, at least not any use that would tap into its true potential.

Figure 6. Text sample from the Wudeng Huiyuan (the same text as in figure 5).

- 14 The screenshot in [figure 6](#) shows the beginning of the body of the text. What I was really proud of and always used as my sales pitch for TEI was the use of the `<RS>`⁴ tag and its `@KEY` attribute. This text recounts the sayings and doings of about two thousand Zen masters, all of them referred to in the text as “師,” which means “the master.” Using the `<RS>` tag allowed me to assign a key to each of these masters and thus distinguish the utterances, find the sayings of specific masters, and much more. Of course, the software to actually do this was not available then, but the concept was still appealing.

4. CBETA

- 15 PhD in hand, I went to Taiwan in February 1998 to attend the founding meeting of the Chinese Buddhist Electronic Text Association (CBETA), and then immediately decided to move there with my family. This provided me with an opportunity to work with a team of very dedicated people on a new digital version of the Chinese Buddhist canon. At that time some lay Buddhists and some researchers in Buddhist studies had started to type texts into their computers and made them available on the internet, then in its infancy, thereby allowing free sharing of these texts. As can be

expected, there was a lack of authoritative editions and scholarly rigor; most texts did not indicate the source of a digital transcription and there was no way to easily compare one to a printed version to make sure it was accurate. CBETA set a goal of providing a digital version of high quality, one that could be useful to both scholars and believers, based for the first batch of common texts on the most authoritative and widely used edition of the Chinese Buddhist canon, the *Taisho Shinshu Daizokyo* compiled in Japan during the 1920s and 1930s.

- 16 Such a project could not be undertaken without a carefully designed workflow and a well-defined text format. In my view, the only possible candidate for this was TEI. Nobody else on the team had ever heard of it, but fortunately they accepted my suggestion and since then the master text format (figure 7 shows an excerpt from such a text) and a number of published versions derived from it have all been encoded in TEI, with the most recent version in TEI P5 and Unicode.⁵

Figure 7. Text from the Taisho Tripitaka in the first XML version, ca. 1998.

```

<type="orig">((沙門...圖))二十四字=((三藏法師義淨奉詔譯))九字【宋】【元】【宮】，((唐三
藏法師義淨奉詔撰))十字【明】</note><app word-count="24" n="0001003"><lem
wit="【大】">沙門義淨從西國還在南海室
<lb n="0001a07" ed="T"/>利佛逝撰寄歸并那爛陀寺圖</lem><rdg resp="Taisho"
wit="【宋】【元】【宮】">三藏法師義淨奉詔譯</rdg><rdg resp="Taisho"
wit="【明】">唐三藏法師義淨奉詔撰</rdg></app></byline>
<lb n="0001a08" ed="T"/><p id="pT51p0001a0801">觀夫自古神州之地。輕生<note
n="0001004" resp="Taisho" place="foot text"
type="orig">殉=徇【三】【宮】*</note><note n="0001004" resp="CBETA"
type="mod"><todo type="*" />殉=徇【宋】【元】【明】【宮】*</note><app
n="0001004"><lem wit="【大】">徇</lem><rdg resp="Taisho"
wit="【宋】【元】【明】【宮】">徇</rdg></app>法之賓。顯法師
<lb n="0001a09" ed="T"/>則創闢荒途。獎法師乃中開<note n="0001005" resp="Taisho"
place="foot text" type="orig">王=正【明】</note><app n="0001005"><lem
wit="【大】">王</lem><rdg resp="Taisho" wit="【明】">正</rdg></app>路。其間或西
<lb n="0001a10" ed="T"/>越紫塞而孤征。或南渡滄溟以單逝。莫不咸
<lb n="0001a11" ed="T"/>思聖跡罄五體而歸禮。俱懷旋踵報四恩以
<lb n="0001a12" ed="T"/>流望。然而勝途多難寶處彌長。苗秀盈十而
<lb n="0001a13" ed="T"/>蓋多。結實罕一而全少。寔由茫茫象蹟長川
<lb n="0001a14" ed="T"/>吐赫日之光。浩浩鯨波巨<CB05145>起滔天之浪。獨
<lb n="0001a15" ed="T"/>步端門之外。<CB00010>草薺而投身。孤<note n="0001006"
B:--@f51n2066.xml 4% (198,96) CVS-1.22 (nXML Invalid Wrap) Fri Sep 7 3:51P
<!ENTITY CB05145 "<gaiji uniflag='0' cb='CB05145' des='[堅-臣+(虎-七+(一/谷))]'
nor='壑' />" >
<!ENTITY CB05418 "<gaiji uniflag='0' cb='CB05418' des='[虔-文+攵]' uni='458D'
nor='虔' moikvo='M075330' mofont='Moiikvo_M114' mochar='7530' />" >
B:--@f51n2066.ent 86% (54,94) CVS-1.16 (nXML Invalid Wrap) Fri Sep 7 3:51P

```


5. TEI Work

5.1 Character Encoding Workgroup

- 17 Around this time (1999, to be precise), John Unsworth, Claus Huitfeld, Lou Burnard, and others were trying to reactivate the TEI project, which had lost a bit of steam since the publication of P3 in May 1994 (Sperberg-McQueen and Burnard 1994). John was convinced that a consortium would be best suited for this, and that it should in some ways be modeled on the Unicode Consortium, the body that maintains the Unicode Standard together with the ISO (International Organization for Standardization). After a long period of preparation, the TEI Consortium was officially incorporated in December 2000 and development and maintenance activity of the TEI Guidelines could resume. The first members' meeting, convened by the late Antonio Zampolli (1937–2003), was held in Pisa, Italy, in November 2001, and I was elected a member of the newly formed Technical Council.
- 18 A long list of things urgently needed to be addressed, including the XML-compatible version of the Guidelines, which was eventually published as P4 in 2002, and a complete overhaul of the way character encoding was addressed within the TEI. For the latter undertaking, I was tasked to form a working group.⁶ The TEI Character Encoding Workgroup's efforts resulted, among other things, in the so-called Gaiji module of P5,⁷ which allows users of the TEI Guidelines to encode characters within their texts which have not yet been added to Unicode. As the name suggests, the creation of this module was initially driven by the desire to address the problem of widely varying orthography in texts printed in the premodern era across East Asia, but it proved applicable to many other cases as well.
- 19 Beyond the direct realm of character encoding, the proposal made by this working group also had an impact on the whole underlying text model of the TEI. If markup can be used for the task of establishing the characters used to represent the text, in addition to describing features of the text, which had been the assumption so far, then this markup-containing text can in principle also appear in some values of attributes, such as labels, representations for a lemma, as well as on <orig> or <sic>, to give just a few examples. All of these instances of attributes that could now

contain markup had to be reformulated as elements, where the value could be textual content, so that the same feature could be expressed without using attributes, and this in turn paved the way to creating elements such as <choice>, which describe different logical paths through a document.

- 20 The old assumption underlying the early work of the TEI (at least in some people's minds), that markup is what is in angled brackets and if one takes that away, the pure, plain text would be left over for the reader to inspect (which had not been entirely true anyway), completely fell apart. Even the generation of such plain text now involves proper parsing and processing of the text with the XML toolchain, not just a few regular expressions that a Perl programmer might come up with.

5.2 Council: P5

- 21 After sitting on the TEI Technical Council since 2001, to my surprise I was appointed chair in 2003. Our task was to invent and implement the major architectural changes that were required for P5, including a whole new templating infrastructure based on what came to be known as ODD.⁸ Fortunately, the never-tiring Sebastian Rahtz (1955–2016) was a core member who did the lion's share of the work, along with Lou Burnard, James Cummings, and Laurent Romary, so my job was simply to keep the ball rolling, schedule meetings, and the like. Nevertheless, when we actually managed to publish P5 on schedule at the TEI meeting in Maryland in 2007, it was a great relief and allowed me to shift focus to other work.

6. *Daozang jiyao*

- 22 Sometime around 2005, we started a project at our institute in Kyoto, led by the late Monica Esposito (1962–2011), which aimed at research in Daoism during the Qing period, focusing on one of the major text collections of that period, the *Daozang jiyao* (DZJY). As happens frequently with premodern texts and text collections in China, the content is somewhat fluid: after the initial printing, there have been additions, reorderings, and other changes to the overall appearance, as well as a reprint at the beginning of the twentieth century, which again was fluid in its composition. In this project, a digital edition was to re-create this complex history and allow researchers to inspect the different versions, compare the content, and trace the development of texts. That was the plan.

- 23 After more than ten years of working with markup and digital texts, it was obvious to me that this had to be done in TEI. I started to train our small team of researchers with a pilot project while designing the workflow. Some of the digital transcriptions were acquired from a company specializing in the production of text databases, others were keyed in by our partners in Taiwan or China, and some were produced in-house. In any case, the texts came in a plain-text format, where every line in the text file represented a line of the original woodblock print. We converted these text files to structured XML files, which were then proofread and enhanced with information about textual variants. Or such was the plan. Then reality intervened.
- 24 As it turned out, mixing the two very distinct tasks of proofreading the digital text against its original source and enhancing this digital text with readings from other sources was not making the work easier, but rather required a constant switching of contexts and became a burden that slowed down progress. On top of that, translating the observations our researchers made on the texts into the correct markup constructs also required much more training than expected and was undertaken less than unenthusiastically by those involved.
- 25 This caused intense discussions among members of the team and we started to look in different directions for solutions. At that time, I also started to dig further into editorial theory, scholarly editing, and the way these fields were changed by the introduction of digital tools. In the end, we decided to separate these concerns and created separate files for each of the versions we edited—both a digital facsimile and a documentary transcription—and separately a new edition, which also included punctuation and standardized orthography that reflected the editorial views of the team.
- 26 The work on the *Daozang jiyao* thus proved to be a good opportunity to further refine both the theoretical approach and the practical implementation of working with premodern Chinese digital texts. Interestingly, at the same time, the TEI community was updating and refining the TEI textual model,⁹ which used to consist of “the one and true” digital representation of the text in the form of a transcribed text with markup applied to it, to which text model they added the notion of a digital facsimile as a separate representation of the text, which shortly afterward was followed by a way to document the physical representation of the text, via `<sourceDoc>`. This development shows how the practice of working with digital text and markup evolves over time and improves our understanding of both the textual features and how markup should be done, a fact that is also documented on an almost daily basis in the discussions on the TEI mailing list.

7. Kanseki Repository

- 27 By the year 2010, the practice of using separate text files for different witnesses of a text had become well established in our workflow. For tracking changes to these files, we had used version control tools from the start. At some point, we realized that the modern distributed variety of these tools, Git and GitHub, not only had the potential to solve the problem of keeping track of changes made to a file, but could also be used to hold all witnesses of a text in one repository, each of them represented as a “branch.” (In the terminology of version control software, a branch is one current state in the editing history of the file, which has been given a name to make it easy to address it and to track changes along a specific trajectory.)
- 28 The distributed nature of this toolchain, which unlike earlier version control systems does not require a central authority, also seemed to have the potential to solve another problem I had been trying to solve almost from the beginning of my work with digital texts. As stated already, one of the aims of my work from the outset was to make a digital version of a text at least as versatile as a printed scholarly edition. For me, this also included taking ownership of one specific copy of such an edition and tracking the work by adding marginal notes, comments, and references directly into the book. With GitHub as a repository for texts and Git as a means to control the various maintenance tasks, researchers interested in a text could clone the text, add their own marginal notes, then make their version of the text available to us or any other researcher to integrate, if we so chose.
- 29 A Git workflow can use any kind of digital material, but it works better with textual material as opposed to images or videos, and even better for texts that use lines as a structural element. This again is where the plain text we used in the *Daozang jiyao* project worked better than did the XML tree structure, which is at the core of every TEI file.
- 30 When I first presented this idea at the TEI conference in Würzburg in October 2011, I got this comment via a tweet from one of the most respected members of the TEI community (figure 8: @rahtz: interesting that @cwittern thinks <> is hard, Git is easy. #tei2011).

Figure 8. Aurélien Berra retweeting Sebastian Rahtz's tweet.

- 31 As described in that talk (published as [Wittern 2013](#)), the text format used here is not simply plain text, but rather an extended form of the text format used in the Emacs [Orgmode](#),¹⁰ in spirit comparable to the much more frequently seen Markdown, but better. The defining difference here is the more elegant and functional choice of markup elements, and the fact that the format was originally conceived as the base for a note-taking and scheduling application, so the markup itself and the software that operates on it are essentially one unit, and the development of the software (which is itself community driven) informs the choices and considerations for markup constructs. For the DZJY project, we added a few more conventions, to accommodate our specific needs, but without changing any of the essential features. Org mode uses what I called an “implicit markup,” which is exactly the opposite of XML. Org mode’s markup is as short as possible and in many cases derived from context. An asterisk * followed by a space at the start of a line indicates a heading of level one, instead of TEI’s `<div>` followed by a `<head>`¹¹ (and the corresponding closing tags to convey this information).
- 32 From the beginning, the DZJY was in my view itself a pilot project for a much larger project, on which preparatory work started in earnest in 2012: the [Kanseki Repository](#) (GitHub username [@kanripo](#)).¹² *Kanseki* here is the Japanese term for premodern Chinese texts, and “repository” to be understood not in the technical sense, but more in the sense used by scholarly archives and libraries.

- 33 Building on the experience of the DZJY, the Kanriipo project sought a firm theoretical foundation for the creation of digital textual artifacts, based mostly on the German tradition of scholarly editing and its distinction between “documentary edition” and “interpretative edition.” These two types are distinguished through naming conventions for the Git branches. Documentary editions are also represented through digital facsimiles, which can be called up to be displayed side by side with the transcribed text. Interpretative editions may normalize the characters used to modern forms, add punctuation, and also make it possible to add translations and semantic annotations.
- 34 From earlier textual projects, such as ZenBase, CBETA, and DZJY, but also from other sources available on the Internet, we have compiled an initial catalog of about 10,000 titles to be included in a first phase of the project; this catalog is also being supplemented by users who deposit whatever texts they are interested in into the repository. Since the initial publication on GitHub in September 2015, and the launch of a dedicated website in March 2016, usage has been increasing slowly but steadily.

7.1 Kanriipo Project Details

- 35 All the texts are freely available on GitHub in their source form. This repository of texts can be accessed through the kanriipo.org website, but also through a module of the Emacs editor called Mandoku. This allows users to query, access, clone, edit, and push the texts directly from their own computer. Reading, commenting, and editing do not require internet access.
- 36 While not yet a full realization of the original vision, this project is currently the best compromise I know of between allowing the researcher (user) to take complete ownership of a text—not just in the technical sense, but also in a practical sense of being in a position to actually be able to edit the text in a way that is meaningful in the context of their aims—and authoritative vetting and editorial quality assurance.
- 37 Figures 9, 10, 11, and 12 demonstrate the concept and functions of the Kanseki Repository. On the website, users can search for texts or browse the catalog. Once a text is found, the webserver reads it from the GitHub repository and serves it to the user. For most texts, there are different editions to choose from; usually both documentary and interpretative versions exist. For many texts, there is

also a digital facsimile, which can be called up alongside the text; if there is more than one edition documented with a digital facsimile, the others can also be directly inspected on the page for the text on the Kanseki Repository website.

Figure 9. A text in the Kanseki Repository.

- 38 In the screenshot in figure 9, there is a link at the top of the page labeled “GitHub,” from which the source of the text can be directly accessed. A user who wishes to make changes to the text, by correcting, annotating, or even translating it, can transfer a copy of this text from the public @kanripo account, either by cloning it to their own account on GitHub, or by downloading it locally.
- 39 The user can also log in to the Kanripo website with their Github credentials. When this is done for the first time, the user has to grant the Kanseki Repository access to their repositories. In addition, a new repository “KR-Workspace” is created; some settings related to the use of the Kanseki Repository are stored here. (Most websites store this kind of information in their own

database, with no direct access to it for the user. KR does it in this way (1) to allow the user control over their data and (2) so that the user's preferences and settings can be applied to different applications with which the user might access the KR.) When the user selects a text for display that they had previously cloned to their own account, the text shown will be their own private version, with all changes and annotations, not the public one from @kanripo. See [figure 12](#) for an example. Other customizations and options become available once logged in.

7.2 KR-Shadow

- 40 What I have shown so far is for users interested in exploration or close reading. For distant reading, text analysis, and similar purposes, a separate account [@kr-shadow](#)¹³ has been created on Github. You will find here the texts of the “master” branch, which is usually the normalized and edited version of the text in a form that makes it easy to download the whole archive at once.

7.3 Mandoku

- 41 As mentioned, the texts can also be accessed from the text editor Emacs, which is available on all major platforms. This is intended for people who work intensely with a text, for example as the topic for a PhD thesis. The Emacs module [Mandoku](#)¹⁴ provides ways to search the KR, clone texts, create new branches, and many other functions. All other Emacs extensions and modules can also be used. [Figure 10](#) shows an example of a text with its digital facsimile, and [figure 11](#) shows the same poems, rearranged by line, with a translation added. In the middle there is an example of an inline note. And finally, [figure 12](#) shows the same text, pushed to the user's account and displayed from there on the Kanripo website.

Figure 10. A text from the Kanseki Repository, side by side with a facsimile, displayed using the Emacs module Mandoku.

Figure 11. The text with translation, also showing an annotation about a Chinese term.

Figure 12. The text with translation, now pulled from the user's GitHub account.

The screenshot shows a GitHub repository page. The header is red with white text: 'KR6q0332 石屋清洪禪師語錄--(參學門人)至柔 (trans-en)'. Below the header, there are two main sections: 'Versions' and 'Contents'. The 'Versions' section lists three versions: '[master]', '[電子佛典集成]', and '[已讀]'. The 'Contents' section lists two items: 'No.1399-A 福源石屋洪禪師語錄原序' and '山居詩'. The '山居詩' section is expanded, showing a table with three columns: the original Chinese text, a Pinyin transcription, and an English translation. The table contains several lines of text, including '余山林多暇。', '曉睡之餘。', '偶成偈語自娛。', '紙墨少便不欲紀之。', '雲衲禪人請書。', '雲衲 [yúnnà]', '蓋欲知我山中趣向。', '於是靜思隨意走筆。', '不覺盈帙。', '故掩而歸之。', '復囑慎勿以此為歌詠之助。', '當須參意。', and '則有激焉。'. The English translation for the first line is 'In the mountains and woods, I am always free.' and for the last line is 'only then will they be able to move you.'

7.4 Relationship to TEI

- 42 I see the approach described here not as an alternative to TEI but as a useful extension. The original concept for the Kanseki Repository did include a transformation tool that would take the different versions of one text and wrap them together into one XML file. This is a “simple” transformation, but there has so far been no opportunity to actually implement it. The other direction, from TEI P5 to a repository of a text with witnesses as branches, does exist, but only for internal use, as it requires a lot of assumptions about source and target. Seen from this angle, the KR can also be envisioned as a postproduction tool, which opens up a new publication avenue for the texts and for users’ engagement with them.

8. In Parting

- 43 As shown in the above examples of projects I have been involved with over almost thirty years, we have come a long way in sophistication of textual awareness and technical means to model it. The TEI conferences, as well as job advertisements for early career academic positions, for example, show that there is a thriving community of experts who engage in a high-level discourse that further advances our understanding of complex features of texts and ways to represent such

features in machine-readable form, readily available for display and analysis. This community is building a bridge right over the deep ridge between researchers in humanities on the one side and computer and technical science on the other.

- 44 For me, the question still remains: how can we make sure that the fruits of our efforts, the sophisticated digital editions of texts, published online in most cases, actually reach the researchers that want to make use of them? And do they reach them in a way that is most beneficial to them? For example, a PhD student working on the development of medical terminology in premodern China, or a researcher interested in changing concepts of relationships as reflected in British novels of the eighteenth and nineteenth centuries—for this kind of research, direct access to the source files of texts, and ideally also ways to select, group, and annotate such files, would be necessary, as well as knowledge on how to filter out exactly what is needed for the question at hand and discard the rest.
- 45 Do we envision such work to be done by the researcher themselves? Or does this require a team of specialists from different backgrounds? To what degree should we open the black box which a web-based critical edition is to many of its users, and allow them (in fact force them) to look beyond the surface?
- 46 As might be obvious by now, to me the answer to this question is: yes, we need to train young researchers in humanities fields to have a minimum understanding of text encoding, text processing, and the underlying technologies. This should be a general requirement, not just an optional minor or a postgraduate-level course. And in an ideal world, I think I would like to introduce them to Emacs, or to a similar tool that allows them to take control of their digital life, rather than being limited to pressing a few colorful icons.

BIBLIOGRAPHY

- App, Urs. 1993. "Guidelines for the Creation of Large Chinese Text Databases." *Electronic Bodhidharma* 3.
- . 1995a. "The Code of the Codex." *Electronic Bodhidharma* 4.
- . 1995b. "Dog Ears and SGML." *Electronic Bodhidharma* 4.

- . 1995c. “From Dumb to Intelligent Text.” *Electronic Bodhidharma* 4.
- App, Urs, Christian Wittern, and Kumiko Fujimoto. 1995. *ZenBase CD 1*. Kyoto, Japan: International Research Institute for Zen Buddhism. CD-ROM.
- Goldfarb, Charles F., and Yuri Rubinsky. 1990. *The SGML Handbook*. Oxford: Clarendon Press.
- Lunde, Ken. 1993. *Understanding Japanese Information Processing*. Sebastopol, CA: O'Reilly.
- . 1999. *CJKV Information Processing*. Beijing; Cambridge [MA]: O'Reilly.
- Lunde, Ken. 2009. *CJKV Information Processing*. 2nd ed. Sebastopol, CA: O'Reilly.
- Sperberg-McQueen, C. M., and Lou Burnard, eds. 1993. Introduction to *Guidelines for Electronic Text Encoding and Interchange*, version 2 (TEI P2), archived document p2ab.p2x. Chicago; Oxford: Text Encoding Initiative. Accessed April 20, 2020. <https://tei-c.org/Vault/GL/teip2.tar.gz>.
- . 1994. *Guidelines for Electronic Text Encoding and Interchange* (TEI P3). 2 vols. Chicago; Oxford: Text Encoding Initiative.
- TEI Consortium 2020. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.0.0. Last updated February 13, 2020. <https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/>.
- Wittern, Christian. 2013. “Beyond TEI: Returning the Text to the Reader.” *Journal of the Text Encoding Initiative* 4. <https://journals.openedition.org/jtei/691>; doi:10.4000/jtei.691.

NOTES

- 1 The standard reference for all East Asian encodings is the work by Ken Lunde, which started out as a textfile called japan.info, circulated via email and on message boards, and later was expanded and published as a book (Lunde 1993). Subsequently, he also integrated information about Chinese, Korean, and eventually Vietnamese encodings (Lunde 1999). Now in its second edition (Lunde 2009), the latter also covers input methods, fonts, printing, and more; it is still a convenient reference for everything relevant to information processing in East Asia.
- 2 See Publications: *Electronic Bodhidharma*, International Research Institute for Zen Buddhism, accessed July 22, 2019, http://iriz.hanazono.ac.jp/frame/book_f5.en.html.
- 3 From the introduction to P2 as of April 1993 (document p2ab.p2x) contained in the TEI Consortium archive, accessed April 20, 2020, at <https://tei-c.org/Vault/GL/teip2.tar.gz>. (Accessed 2020-04-20),

- 4 The same as <rs>, which in TEI P5 is glossed: “(referencing string) contains a general purpose name or referring string.” TEI Consortium 2020, Appendix C: Elements, “<rs>,” <https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/ref-rs.html>.
- 5 The most recent version of the whole textual database is available in the CBETA XML P5 GitHub repository, accessed April 20, 2020, <https://github.com/cbeta-org/xml-p5>.
- 6 TEI Character Encoding Workgroup, last updated September 16, 2007, <https://tei-c.org/Vault/Workgroups/CE/>.
- 7 This is now in TEI Consortium 2020, 5: “Characters, Glyphs, and Writing Modes,” <https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/WD.html>.
- 8 ODD (One Document Does it all) was the nickname for the composition of the Guidelines, inspired by literate programming, which included the description of the outcome both in its human-readable form and in the machine-readable schema. The P5 version of this schema definition was first called the Son of ODD, but later was simply known as ODD; see TEI Consortium 2020, 22: “Documentation Elements,” <https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/TD.html>.
- 9 I am specifically referring here to the work by the Workgroup on Genetic Editions, chaired by Fotis Jannidis, which published a draft in 2009. For details, see “An Encoding Model for Genetic Editions,” accessed July 3, 2020, <https://tei-c.org/Vault/TC/tcw19.html>.
- 10 Accessed May 18, 2020, <https://orgmode.org/>.
- 11 For a full description of this format, see “The Mandoku Text Format,” accessed April 20, 2020, <https://github.com/mandoku/mandoku/blob/master/doc/mandoku-format-en.org>.
- 12 Accessed June 24, 2020, <https://www.kanripo.org/> and <https://www.github.com/kanripo>.
- 13 Accessed June 24, 2020, <https://www.github.com/kr-shadow>.
- 14 Accessed May 18, 2020, <https://github.com/mandoku/mandoku>.

AUTHOR

CHRISTIAN WITTERN

Christian Wittern studied philosophy, German literature, and Sinology in Tübingen, Hamburg, Hangzhou, Kyoto, and Göttingen. During graduate studies in Kyoto, he participated in the Zen Knowledgebase project led by Urs App at Hanazono University, which published digital texts and research tools on a CD-ROM in

1995. After receiving a PhD, he moved to Taiwan to work at the Chung-hwa Institute for Buddhist Studies and with the Chinese Buddhist Electronic Text Association (CBETA), for which he has served as adviser since its foundation in 1998. He currently works at the Center for Informatics in East Asian Studies, Institute for Research in Humanities, Kyoto University. One of his current research projects is the Kanseki Repository, an open-source collection of premodern Chinese texts, currently holding almost ten thousand items.