# Chinese TEI – A guide to using TEI with Chinese texts

This text is:
Marcus Bingenheimer: “Chinese TEI – A guide to using TEI with Chinese texts.”
Published in print as “Bianzhe xu TEI yunyong lue shuo 編者序:TEI 運用略說” in: *TEI shiyong zhinan - yunyong TEI chuli zhongwen wenxian* TEI 使用指南—運用 TEI 處理中文文獻 [Chinese TEI – A guide to using TEI with Chinese texts]. Taipei: Taiwan E-learning and Digital Archive Program 數位典藏與數位學習國家型科技計畫, 2009. ISBN:978-986-01-8092-3.

Author: Marcus Bingenheimer
Autumn-Winter 2008
Revised: May 2009



# 1. Introduction & Acknowledgments

There are many meeting places for the Humanities and Computing Sciences. but few where they actually overlap. Markup technology is one of these. Markup allows us to model scholarly editorial practices directly in the context of digital text and is therefore relevant for all scholars working with digital text. There are a great number of markup standards in many different fields. People use MathML[1] for mathematical formula, the Music Markup Language[2] to express musical notation in digital form, and MIX to store technical information about image files. For digital archives, standards like METS and MODS describe the information necessary to bundle files into composite digital objects and allow librarians to archive and circulate them.
There is, however, one standard that is special in its comprehensiveness, history and usefulness. TEI (Text Encoding Initiative) is the most important achievement in the fledging field of Humanities Computing. It is an open community-developed markup standard for texts of all periods and all languages. In essence TEI allows us to mark up all sorts of information about the text: its sources, structure, variants, omissions, notes, indices, tables, but also our interpretation, our own notes and corrections, even to the point of recording the degree of certainty with which an assertion is made. With

1 http://www.w3.org/Math/
2 http://www.musicmarkup.info/

its more than 500 elements, attributes, classes, and modules, its numerous examples and extensive prose guidelines TEI is probably the most comprehensive markup standard available in the Humanities. For scholarship in the Humanities it has emerged as the *de facto* standard.

The TEI-consortium has recently made efforts to internationalize the standard by providing an infrastructure for partners across the globe to localize the standard in their own languages. Dharma Drum Buddhist College, the Taiwan e-Learning and Digital Archives Program (TELDAP) and the TEI consortium have over the last three years cooperated to localize the TEI standard. The book you are reading is the result of this cooperation.

The purpose of this book is twofold:

1. We want to provide a introduction to TEI markup with special emphasis on problems encountered in Chinese texts. We will discuss basic concepts of a markup project and give concrete examples of how to markup textual phenomena.[3]
2. The book is also intended to serve as a reference work for encoders. It includes the fruit of many years of translating parts of TEI. Included are:
    - TEI-Lite: a short introduction to marking up texts with TEI. We have annotated this introduction freely, especially in cases where the examples were of a kind that needed more specialized knowledge of European literature.
    - The translations of element and attribute definitions. All of these translations are available on the TEI website and will be maintained there by the TEI consortium.[4]
    - Localized and translated examples for each element.
    - Selected translation from the prose guidelines: Chapter 2

Apart from the translations presented here the cooperation between DDBC and the NDAP has yielded a Chinese translation of the ROMA interface (http://www.tei-c.org/Roma/; see. Ch.4 *Start with the schema* for how to use Roma), and a number of workshops introducing TEI at the Academia Sinica by Julia Flanders and Syd Bauman (Spring 2006), Sebastian Rahtz (Spring 2008 and 2009) and myself (Winter 2005-2006 and Autumn 2006).

The people who contributed to the localization of TEI are many. Here I want to mention the Metadata Architecture and Application Team (MAAT) at Academia Sinica, who did a first draft translation of TEI Lite P4. Professor Fu Xinjia 傅心家 from Qinghua University, who organized a workshop with Julia Flanders and Syd Bauman, and Professors emeriti Xie Qingjun 謝清俊 and Xie Yingchun 謝嬴春 , who were among the first to support TEI in Taiwan. Christian Wittern was the first to introduce me to TEI and always took time to answer questions. My friend and colleague Prof. Aming Tu 杜正民 has always encouraged me to keep working on this and other TEI related projects and provided his unflagging support. Without him this localization would not have happened.
The bulk of the at times extremely difficult technical material was translated and retranslated by Huang Weining 黃韋寧 and Wu Tianling 吳恬綾 (element and attribute definitions), Liu Chunyou 劉純佑 (Guidelines Chapters 2. and 5.) and of course Xie Xiaolin 謝筱琳 (localization of examples and administration). All of them have worked on the TEI Lite P5 translation as well. Rong Xiqin 戎錫琴 translated this introduction. The way they came to terms with the difficult technical matter of the guidelines is outstanding and deserves highest praise. The Guidelines, though well written, often force

3 Many of the examples are taken from the workshops I taught at Academia Sinica 2005 to 2007. The handouts and slideshows for these and other TEI workshops in Taiwan are available at: http://buddhistinformatics.ddbc.edu.tw/~mb/webclassmb/teiWorkshop/indexTei.html.

4 As always in the case where technical standards are translated, the version of the original is authoritative, i.e. in case of doubt follow the English.

even native speakers of English to reread passages several times and their localization called for excellent linguistic competence as well as creativity and critical thinking. On the administrative front, Chen Meizhi 陳美智 and her team at Academia Sinica have done a great job of organizing workshops and helping with the budget and organization of the first and second print run.
With all these wonderful people involved it is clear that that I alone am to blame for any remaining mistakes. Indeed, I do not only take responsibility for them, but also invite the reader to communicate them to me so that we can improve this localization in the future.

I hope this localization of an important standard in Humanities Computing will help those who work with digital Chinese texts to integrate their efforts with the international mainstream in this area. The textual universe of Chinese culture is a fascinating part of mankind's cultural inheritance. Open standards like TEI guarantee that the treasures of Chinese history can be made freely accessible for research and enjoyment far into the future. Therefore, in the spirit of sharing, this effort is dedicated to all friends and admirers of Chinese culture.

## 2. A short history of TEI

Most of the markup languages mentioned above use a common syntax – XML (eXtensible Markup Language). Since 1998, the year XML appeared, this syntax has proved useful for the design of many interchange standards. The concept of markup has been around before and the first attempt to design a syntax for markup standards was SGML (Standard Generalized Markup Language)[5]. Designed by Charles Goldfarb and others, SGML was an intelligent syntax and in some ways more comprehensive and flexible than XML. Unfortunately, it proved difficult to write even basic applications, such as validating parsers, for SGML and for many years institutions that had to markup electronic text had no choice but to resort to expensive proprietary solutions. With the advent of XML the situation improved – due to its tighter and better circumscribed syntax it became possible for a single programmer to develop tools that allow the validation and transformation of a text encoded in XML.[6]

TEI was conceived in 1987 when an international group of specialists in electronic text met at Vassar College in Poughkeepsie, New York. They decided to develop a standard for the markup of electronic texts and agreed on general design principles. A first draft (P (=Proposal) 1) was completed in 1990 by Lou Burnard and C. M. Sperberg-McQueen. P2 appeared 1992, and with P3 in 1994 TEI achieved a degree of stability and maturity that served it well for a number of years.
During the 1990s TEI was supported by grants from the US National Endowment for the Humanities, the European Union, the Canadian Humanities and Social Science Rsearch Council, and the Mellon Foundation, and by the institutions who had served as hosts during the development of the standard.[7] Ten years later it was felt that TEI needed a more formal organizational structure and in December 2000 the TEI consortium was founded. The first membership meeting was held in Pisa in 2001. Since then membership meetings have been held every year in various locations and the TEI community has been steadily growing. The consortium is governed through its bylaws and managed by an elected board of directors and a technical council. At the time of writing, the consortium has more than 80 members, almost all of them in Europe and North America. There are two member institutions from

5 SGML became an ISO standard (ISO 8879) in 1986.
6 One of the better open-source XML-Editors available that was basically developed singlehandedly is the XML Copy Editor (http://xml-copy-editor.sourceforge.net/).
7 Currently TEI is hosted at the universities of Oxford, Brown, Virginia and Nancy.

Taiwan, but, to my knowledge, none in Japan, Korea or mainland China.
After an organizational structure for TEI had been created technological developments necessitated an update of the standard. P3 was still expressed as an SGML standard. That meant the texts encoded in P3 were to be validated with DTDs and highly customized tool chains were typically needed to manage them. With the appearance of XML in 1998 it became clear that the TEI standard had to be made compatible with this new format. In 2002 version P4, edited by Lou Burnard and Syd Bauman, realized TEI in XML. As technologies go, XML turned out to be quite successful and gave rise to a number of new technologies – namespaces, schema-languages, XPath, XSLT, XQuery, and a host of other X-standards (some more successful than others). These new technologies changed the way digital text is produced, managed, interchanged and archived. The TEI consortium rose to the challenge and continued to keep its standard in sync with the technological progress in the wider field of computing. The result of its effort is P5, which was released in November 2007 and which introduces a number of important improvements including namespace-awareness and a more robust customization mechanism (ODD). Especially interesting for Chinese digital text is the addition of the *gaiji* module that allows for the description and management of characters missing from the standard XML character-set.
Currently it is easy to see whether a document conforms to P5 or to a previous version: The root element of a TEI P5 document is <TEI>, and the document must be in the TEI namespace at http://www.tei-c.org/ns/1.0 . For all earlier versions , the root element is called <TEI.2>.

## 3. What markup does for you and what else you need to manage digital text

If you are reading this, chances are that you are dealing with digital text in one form or an other. This section will try to answer the question what TEI can do for you, what it can not do, and what else you might need next to TEI in order to manage digital text efficiently.

TEI allows editors, scholars, and librarians to encode all kinds of phenomena in a digital text that they deem relevant. These could be: aspects of the text itself like its chapter structure; aspects of the source from which the text was created, like a page number; editorial inventions like the addition of a lost passage from an other work etc.

Why should one invest the time to study and use an international standard to do this? Why not just mark texts somehow in any way convenient for the encoder or programmer?
Here are the reasons:

1. Digital text does not live on our own computers only. International standards like TEI guarantee some degree of **interoperability**. The texts we produce will probably be archived in very distant places and times. When a digital Chinese text includes a TEI header a librarian in North America, for instance, has all the necessary information to integrate this digital object into an archive without having to know Chinese.
2. Validation is an essential part of XML based projects. TEI lets users build intelligent, highly **customized validation scenarios**, which can easily be expanded and modified. The ability to customize the schema effortlessly is an important asset for every digitization project.
3. The TEI community provides a growing repertoire of **free software** that saves time and labor. There are freely available stylesheets to transform TEI texts into HTML, PDF, ODT and (perhaps somewhat surprisingly) even to DOCX. There are various customizations of TEI for different user communities, such as EPIDOC for epigraphic material, or TEI Tite for

outsourcing digitization to vendors.

4. A growing number of **XML editors** come **with TEI support** out of the box (Oxygen, XML Copy Editor etc.).
5. Home-made solutions generally lack public documentation, while every aspect of TEI is **amply documented**. More and more of this documentation, like the book you are reading now, is available in different languages.
6. The TEI consortium with its wiki, mailing-lists, special interest groups, and workshops offers the opportunity to participate in a **strong user community**. Here we can learn from the mistakes of others and share information about successful solutions. The rapid technological changes of the last ten years have shown, how important it is to stay in touch with what is happening in the general field of Humanities Computing where new solutions appear as quickly as new challenges.

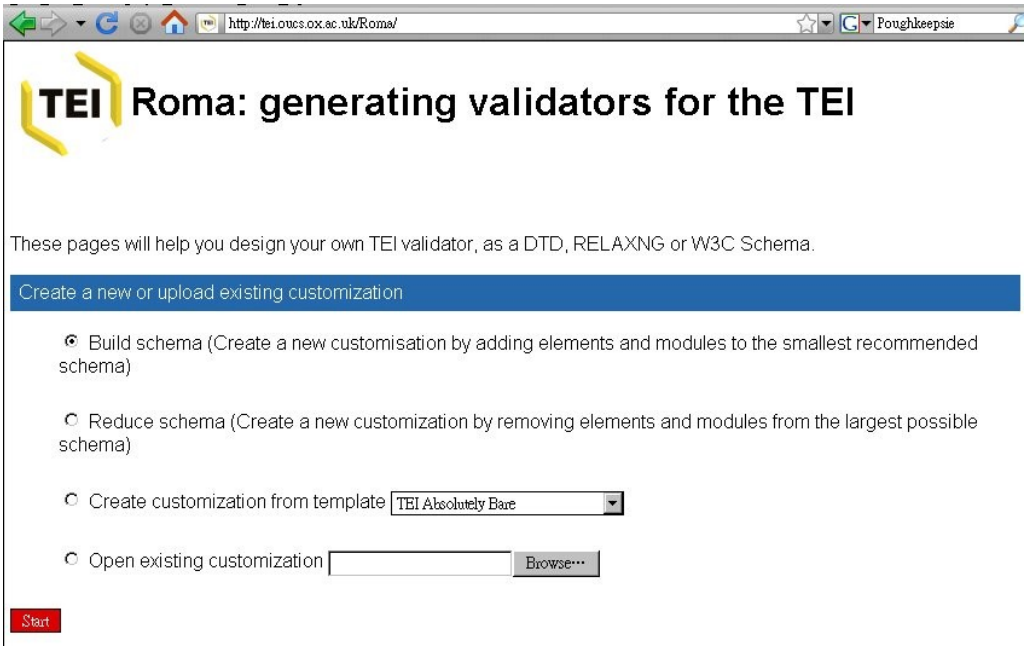
In any organization that maintains a digital archive a large number of skills and standards are needed to manage digital objects. TEI is first of all a markup standard for written text, which was first conceived of before the Internet (let alone digital multi-media content) became widespread. TEI is well placed to connect text to images, audio recordings and even video, but it can not well express the technical metadata for digital multi-media objects. For these dedicated standards such as MIX should be used and then tied to the TEI transcript via wrapper formats such as METS. A digital archive will in all likelihood have to use more than one XML standard to manage its collections.
To manage digital objects and their metadata a fair number of computing skills are needed beyond markup. Digital archives have to administer their own servers and databases as well as develop user interfaces for their collections. If the bulk of data and metadata is in XML a good knowledge of dedicated XML technologies will make its management much easier and more productive, even though some of these are not yet widely taught or used in Taiwan. XML data such as a TEI document can in principle be accessed with any programming language, however, the richer the markup, the more complex is the information that is expressed in it and the more useful are XML technologies for managing it. Adding XSLT and XQuery to a workflow will often simplify the procedure and increase efficiency. It is also a good idea to deploy native XML databases such as eXist, or MarkLogic when dealing with large document collections. In our archive we use XML technologies alongside more traditional relational database systems. Where the data structure is fairly simple, relational databases deliver more speed and more mature tools. Digitizing cultural content, however, rarely results in simple data structures.

## 4. Start with the Schema

In courses introducing TEI the creation and maintenance of the document model or schema is often discussed last. However, there are reasons to put a discussion of the way schemas are created in TEI at the beginning. After all, the ability to validate a document instance against a schema is one of the main reasons to use a XML standard in the first place. It is here, in its mechanisms to manage schemas, that TEI goes far beyond other standards.
A XML standard defines a number of elements and attributes and decides on how these can be used in the document. This information is laid down in the document model or schema, which is written in a formal language. While in the past most schemas were written in DTD syntax, today it is advisable to use languages like XML Schema, Relax NG, or Schematron which are much more expressive than

DTD.[8] Schemas are an indispensable part of every markup project. The encoders must use a schema to validate their documents and see if they conform to the standard. Now, with an average XML standard, say NewsML, users are not usually encouraged to change the schema. You either conform to it or not. What makes TEI special, is that it allows the user to customize the TEI schema in a documented way. You can add and delete elements and attributes, change the content model of elements, restrict element content and attribute values with regular expressions, and more. The changes to the schema are documented in a customization file called ODD file (One Document Does it all)[9]. The ODD file is of course itself a XML/TEI document. From the ODD file a schema can be generated in any syntax.[10] To do so painlessly, the TEI consortium maintains a fabulous tool called Roma (http://www.tei-c.org/Roma/). Roma was written by Arno Mittelbach, and is currently maintained by Sebastian Rahtz at Oxford University.
The first time a visitor comes to Roma he or she is likely to generate a schema from scratch.

Screenshot 1 + 2:




1: Choose "Reduce Schema" to start a new project for which you need a schema which you might want to customize and click "Start". If you use "Reduce Schema" all elements and attribute of TEI will be included in your schema by default. It means you will have the greatest possible selection, but it is not necessary the best choice. A schema should be just right, i.e. contain only those elements that are actually used. The whole idea of schemas after all is is to constrain our choices in intelligent ways to produce a clear and consistent markup. "Reduce Schema" is an convenient choice if you are not yet familiar with the module structure of TEI.[11] More experienced users might want to start with "Build Schema". This way they begin with the TEI schema that contains the fewest possible modules (*tei*, *core*, *header* and *textstructure*).

8 The default schema syntax for TEI is Relax NG. Most customizations can however be expressed in DTD syntax or XML Schema as well.
9 A detailed description of the ODD specification is found in Chapter 22 of the TEI guidelines.
10 Provided all distinctions in the schema are at all expressible in a particular syntax.
11 TEI groups elements in thematic modules. All elements that can appear in the TEI header for instance are contained in the *header* module. This simplifies customizing and managing the schema.


2: Choose a title for your schema and a file name. Do not include whitespace or Chinese characters in the file name. Then select the language for the Roma interface (and don't forget to click "Save", as I usually do).

Screenshot 3:


3: Go to "Modules 模組". The more than 500 elements of TEI are grouped thematically in modules. The modules roughly map to chapters in the prose guidelines. For instance, elements used for in-depth encoding of names and dates are grouped in the module called *namesdates*, which in turn is the topic of Chapter 15 in the guidelines (*Names, Dates, Peoples and Places*). On the Modules page of Roma you should delete those modules which you don't use (if you started with "Reduce Schema") or add those modules that contain elements you need (if you started with "Build Schema").

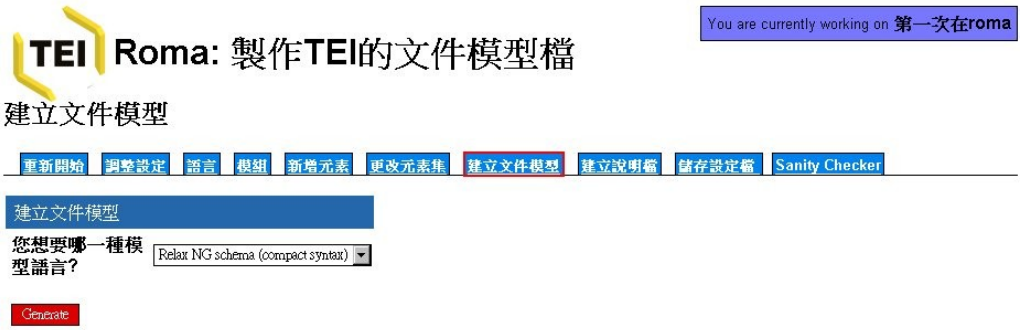
If you are new to TEI the names of the modules will not mean much to you and it will obviously be a problem for you to decide what to keep and what to ditch. This is a catch-22 problem: you can' t determine what you are going to need for a project without being somewhat familiar with TEI, on the other hand you won't get familiar with it if you don't start somewhere.

For first time users it is OK to start with the largest possible schema (coming in through "Reduce Schema"), but you should be aware that the aim is to eventually reduce the number of elements. Ideally, your schema should contain only those modules you actually use in your markup. If you want to see what elements a module contains just click on the module name.[12]

Marking up Chinese material you usually want to include the *gaiji* module that lets you describe annotate existing or define new characters. I also found that I need *linking* and *namesdates* in most of my projects.

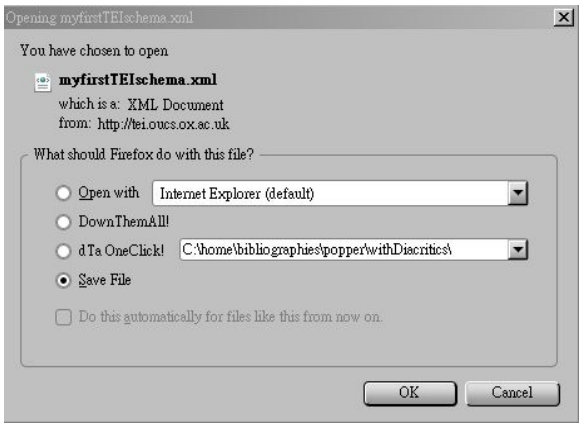
Screenshot 4:

12 Clicking on the module name takes you to an interface where you can further prune each module by deleting single elements. *Core* especially is a large module which might need trimming. *Core* is meant to provide basic markup mechanisms for a large number of things and therefore almost certainly contains elements which you will not need. Take the time to weed out the elements you do not need, you can always add them back again later. The resulting schema will be leaner and easier to manage.




4: Go to "Schema 建立文件模型" Now you are ready to order your schema. Choose the syntax you are using. TEI itself is (for various good reasons) expressed in Relax NG by default, but thanks to the TRANG converter by James Clark this can be converted into XML Schema and DTD syntax within the expressive limits of these languages. If you want to use Relax NG remember that Relax NG comes in two formats: XML and Compact syntax. If you plan to validate your documents on the fly on a server of as part of a processing pipeline, better use the XML format of Relax NG. If validation mostly happens locally, or if you want to look into the schema and see why things work the way they do, you are better off using the Compact syntax. Don't get too concerned about this in the beginning; the syntax you choose makes little difference as each syntax expresses the same abstract document model and will validate your document equally well.

Screenshot 5:




5: Click on "Save Customization" Here you download your ODD specification which describes the schema you just assembled in the last few steps. Save your ODD close to your schema. Remember that it is the ODD file that counts. Though the parser uses the schema you created in Step 4 to validate a document, that schema is just an instantiation of your ODD file. Do not lose your ODD file. You will need it to make further changes in the schema. Roma can always generate a new schema with your ODD file, but it cannot create an ODD file from a schema.

If you need to expand or constrain your schema later, take your ODD file and return to Roma. This time, on arrival you choose "Open existing Customization" and upload your ODD file. Then use the Roma interface to make the changes, save the new schema and ODD, and work with the schema until the next changes are necessary, then the cycle starts anew. Never edit a TEI schema by hand! Chances

are that doing so will to great confusion - technically, mentally and probably financially.
The ODD mechanism allows your project to produce highly customized schemas (e.g. with added elements, restricted attribute values etc.) and still claim TEI conformance. As long as you distribute your texts together with your ODD file, researchers, programmers and librarians elsewhere will be able to handle your content and understand the design principles of your collection.

Screenshot 6:


6. Go to "Documentation". Here you can download a documentation file (as HTML, PDF or a TEI Lite document). The documentation file contains the descriptions and examples for all elements in your schema. For documentation in Chinese chose Chinese at the “Language” Tab. The documentation can be used to train encoders and serve as reference for programmers. It is automatically generated from the guidelines. Therefore print-outs are only useful if you have successfully slimmed down your schema in the "Modules" interface. Otherwise you will end up with descriptions of all 500 elements.

The latest (2008) development in TEI schema management is a standalone desktop version of Roma called Vesta, which provides similar services with no need to be online.

## 5. Build the Header

Every TEI document has the same basic structure:
<TEI>
<teiHeader>.... </teiHeader>
<text>.....</text>
</TEI>

The function of the two main divisions is similar to that of <head> and <body> in HTML. The <teiHeader> contains the metadata concerning the creation of the digital document, such as its sources, editorial conventions, ownership, revision history etc., while <text> contains the document content proper.
Obviously, digital librarians and archivists like the <teiHeader>. It allows them to map bibliographical

metadata to library standards like MARC21 automatically.
The header is an indispensable part of a digitization project with TEI. If the project, as is often the case, consists of many different TEI files, it is sometimes advisable to devise a common stand-off header. The common header, or parts of it, can be referenced from the files containing the <text> via an XML standard called XInclude. Using XInclude ensures that the text can be validated even if header and text proper are in different files. That way the metadata in the header can be maintained more easily.
I often found it helpful for a project team to spend some time on the TEI header before beginning with the actual encoding. It helps to get a clearer idea of the material, division of labor, and what conventions the team is going to use. Many questions regarding the markup only emerge during the actual practice and have to be addressed one by one during the early stages of the project, preparatory work on the header provides the team with a common outlook.

The first level of the TEI Header can contain four sections, of which only the first (<fileDesc> (檔案描述)) is mandatory:

<fileDesc> 檔案描述
<encodingDesc> 編碼描述
<profileDesc> 文件背景描述
<revisionDesc> 修訂描述

<fileDesc> can contain seven child elements: <titleStmt> (題名陳述), <版本陳述>/<editionStmt>, <檔案大小>/<extent>, <出版陳述>/<publicationStmt>, <集叢陳述>/<seriesStmt>, <附註陳述>/<notesStmt> and <來源描述>/<sourceDesc>. Of these only <題名陳述>/<titleStmt>, <出版陳述>/<publicationStmt> and <來源描述>/<sourceDesc> are mandatory.
The smallest possible TEI Header therefore contains only a <fileDesc> which in turn contains <titleStmt>, <publicationStmt> and <sourceDesc>.

Here an example:

```
<teiHeader>
        <fileDesc>
                <titleStmt>
                        <title>七大寺年表 with TEI markup</title>
                        <author>慧珍 (1118-1169)</author>
                        <respStmt><resp>digitized by</resp>
                        <name>杜正民</name></respStmt>
                </titleStmt>
                <publicationStmt>
                        <distributor>法鼓佛教學院</distributor>
                </publicationStmt>
                <sourceDesc>
                        <bibl><title>七大寺年表</title>大日本仏教全書 [Complete Works of Japanese Buddhism].
                100 vols., Suzuki Research Foundation - edition 鈴木学術財団, Tokyo: Kōdansha 講談社, 1972.
                No. 647</bibl>
                </sourceDesc>
        </fileDesc>
</teiHeader>
```

Here we state the title of the digital text,[13] who is responsible for its creation, where the digital version

13 Note that this is not always the title of the original that was digitized. The title of the original can appear in the <sourceDesc>.

is (or was originally) distributed, and where the original text was published.
This minimal header will let you validate the document, but any larger project will want to provide more information about its practices. Remember, that the more comprehensive your metadata is, the better are the chances that your text will survive in future digital environments. Since the Chapter concerning the Header is translated in full below, in this summary I will highlight only a few things which can be done with the header.

Every digitization project has its own idiosyncratic practices – ways to deal with specific problems for which there are no standard solutions. These practices can and should be recorded in the TEI Header. Issues that often crop up when dealing with Western languages are for example the treatment of hyphenation (words hyphenated because of a line-break regularized or not?), or spelling conventions (American or British English?). In a Chinese environment, too, issues regarding normalization or spelling are part of almost every project. Are unusual variants regularized to 通用詞. How are non-Unicode characters realized in the text?
It is important to record these decisions. The place to do so is <editorialDecl> within <encodingDesc>.

```
<encodingDesc>
 <editorialDecl>
        <normalization>
                <p>原文簡體字轉成繁體字</p>
                <p>在日文人名地名保留日文漢字(非改成中文繁                體字)</p>
        </normalization>
        <quotation marks="all" form="std">
                <p>所有的引用號轉成實體(&amp;quot;)</p>
        </quotation>
 </editorialDecl>
 </encodingDesc>
```

For any project where different scripts and/or languages are marked in the text, the language use must be recorded in the <profileDesc>. If needed this record can be quite detailed.

```
 <profileDesc>
 <langUsage>
        <language ident="zh-tw">在台灣講的中文</language>
        <language ident="zh-cn">在大陸說的普通話</language>
        <language ident="en-us">English as used in the United States</language>
        <language ident="en-uk">English as used in the United Kingdom</language>
 </langUsage>
 </profileDesc>
```

This digital document contains only two different languages, but the profile description records four different usages. Wherever possible the respective ISO standards for locations or scripts should be used, only in case where these can not be applied should one use prose descriptions.

There are many more things that can be recorded in the teiHeader, but since we will include a complete translation of the Header Chapter from the guidelines we will not go into further detail here.

# 6. Mark up the text

After you have created a first ODD file, which allows you to order and customize your schema from Roma, and after a first draft for a TEI header with the basic metadata has been agreed on it is time to start with the markup of the text.

**6.1 Structure**

At first you might want to give the text some light structural markup, identifying the major divisions of the text and referencing it back to the source. If you are digitizing a book, for example, you might want to use  (division) for chapters, <p> (paragraph) for print paragraphs and <pb> (page-break) to refer back to the original pagination. If your text is a stele inscription, the whole inscription might be termed a . <p> might not be needed at all, and you would record line-breaks (<lb>) rather than page-breaks (<pb>). Most of the time you will have to take a decision on a few basic questions:

Do you want the structure of your markup to model the structure of a printed or written original, or do you want to structure the text in new ways? That is to say, to what extent is your project a simple digitization, i.e. modeling of a non-digital object and in how far is it a new, independent edition?

Can the encoders add content, for instance in the form of making <note>s and indicating their responsibility?

Is only one source referenced or does the electronic edition combine two or more printed or written sources?

**6.2 Names and Dates**

For many projects a light structural markup might be all that is needed. However, the general trend in recent years is to present electronic text in combination with image data of the source. After the spectacular success of large scale digitization projects like Google Books and Europeana, both of which combine images with electronic text, metadata, and advances in OCR technology, users will more and more demand to consult the sources directly. We want to search large corpora of electronic text, but in the end look at a book page, or at least an image thereof.

This is bad news for text-only archives, which rely on light structural markup to achieve an output that approximates the original. It does, however, reassert the original approach of TEI, that is to offer a highly expressive language to work with digital text, beyond merely mapping the structure. If digital text is seen as a mere digitization of a paper edition, then a digital image and a text file derived from an OCR will in most cases be just as useful. However, by adding "deep" markup it is possible to go far beyond a faithful reproduction of the original. A TEI project should add valuable information to the text that cannot be derived by automatic means.[14]

One of the most general textual phenomena that deserve to be distinguished, indexed and analyzed are names and dates. Most text are about something and things usually have names. The ability to index and connect these is the very basis for more powerful knowledge management. Only if variants of names are identified and connected is it possible to go beyond simple full-text searches -- an important function for classical Chinese text where a person might be referred to by a variety of names (別名, 號, 筆名, 俗性, 法名, 諱名).

The encoder identifies different forms of a name and references them with the *key* attribute to an entry

14 Examples for projects that connect images and text as well as using intensive markup are the Emblem project in Utrecht (http://emblems.let.uu.nl/), *Le mariage sous l'ancien regime* (http://mariage.uvic.ca/) and our own project on Buddhist temple gazetteers (http://dev.ddbc.edu.tw/fosizhi/).

in a authority database or index file. That makes it possible to search for one name variant and retrieve all passages where a person or place is referenced.
We do not have algorithms developed enough to be used to reliably parse a text and identify all names within it.[15] Also, it is not easy to downscale the strategies involving statistical analysis that companies like Google use for searching the Internet to the size of our corpora. On the other hand, scholars would like to search their text copora at least as intelligently as Google searches the Internet. For the time being, human brains are still needed to identify whether two names refer to the same person, and the best way to express such situations in a digital text is by means of markup. Among markup standards TEI is sufficiently expressive to consistently markup a large number of such textual phenomena.

So how do we markup names in TEI? The most general mechanism is the <name> element. You could use the type attribute to distinguish different types of names:

```
<name type="person">張欣怡</name>
<name type="place">金山鄉</name>
```

However, the type attribute should be used with care. Its possible values should be restricted to a list of tokens. This is eminently useful for validation.

In the case of place and person names, moreover, TEI offers dedicated elements. So instead of the above we would usually use:

```
<personName>張欣怡</personName>
<placeName>金山鄉</placeName>
```

TEI offers a lot of elements and attributes to further determine the name. For example you might want to distinguish a persons <forename> and <surname>:

```
<persName>
        <surname>張</surname>
        <forename>欣怡</forename>
</persName>
```

In classical sources names can get quite complex. The founder of the Linqi tradition for instance is in one source referred to 臨濟開法費隱通容禪師. Without modifying the TEI schema every part of this compound could be marked as:

```
<persName key="s26579">
        <name type="lineage">臨濟</name>
        <addName>開法</addName>
        <name type="dharma">
                <name type="common">費隱</name>
                <name type="taboo">通容</name>
        </name>
        <roleName>禪師</roleName>
</persName>
```

15 Attempts to devise ontologies for this purpose have, in spite of considerable efforts in this direction, so far failed to solve problems of scale. With the help of markup ontologies might evolve over time – first in the form of authority databases then as the more complex structures today called ontologies – abstract models of reality.

This kind of extensive markup is of course not always necessary; each project has to decide on the degree of verbosity of its markup in general and on what aspects of a name to mark in particular.

Every project in which names of persons and places are important will maintain an index or an authority database where these are collected and from which they can be referenced. The data format of the index is not really relevant, but if you have only little computing support I recommend you to keep the index in the same format as the rest of the data, i.e. TEI. If your team includes competent programmers, who can connect the XML-data of the source to a relational database that will do as well. In either case it is desirable to build the index in a way that the data is reusable in other projects. This means avoiding project-specific IDs, as well as sufficient documentation about how an index entry is constructed and how it should be edited.[16]
For detailed information about the markup of names see Chapter 13 of the full guidelines. The topic is also shortly treated in the translation of TEI Lite (Chapter 10) below.

### 6.3 Editorial Interventions

Using TEI markup digital editors are able to model all editorial practices that are possible in print . Beyond that they are able to realize those that are specific to the digital realm. Common editorial interventions include:
- marking apparent errors and omissions 省略 and making corrections
- introducing regularizations and recording variants 一般化與異體字
- making additions and deletions

Generally, if you come across something that you would wish to amend, comment on, correct or delete you would want to preserve the original reading together with your changes. TEI expresses this in the <choice> element. If for example you consider that in the passage (<ab>) "諸天適興恚怒。彼鬼遂轉端正正。顏貌勝常" the double 正 is an error you can correct it using the <corr> tag, but you should at the same time preserve the original meaning in <sic> (sic = Latin for "just like this"):

<ab>諸天適興恚怒。彼鬼遂轉端<choice><sic>正正</sic><corr>正</corr></choice>。顏貌勝常</ab>

If the changes you make are regularizations not corrections, i.e. the original is not a wrong but merely a different representation of the same thing, you should use <orig> and <reg> instead:
The sentence (<s>): 施先生付了壹佰万美金! could be encoded:

<s>施先生付了<choice><orig>壹佰万</orig><reg>一百萬</reg></choice>美金!</s>

Or a more classical example from the 黃檗山志:

洎獲<choice><orig>寘</orig><reg>置</reg></choice>之于塔分七粒于琉璃器中瑩然光色

If you want to mark up manuscript material chances are you will come across words and passages which have been deleted or inserted by someone either the original author or a later redactor of the text.

16 It is also good practice to make useful authorities available to others, either by providing web-services or as raw data. For an example see e.g. the authority databases at: http://authority.ddbc.edu.tw/.

Here in a letter head (<head>) the "Dear" has been crossed out:

~~親愛的~~淑情你好 [??image file for this with handwriting??]

In TEI this can be expressed as:

<head><del hand="author">親愛的</del>淑情你好</head>

To mark additions or insertions use <add>:

淑情~~你~~\妳/好

In TEI this becomes:

<head>淑情<del>你</del><add rend="above">妳</add>好</head>

Sometimes a character, a word, or a passage will be illegible due to some reason or other.
For this case TEI offers the <unclear> and <gap> tags.
For instance in the verse line (<l>) 假使虛空裂□地皆振動 the character between 裂 and 地 has been blotted out by an ink stain.
You can express this as:
<l> 假使虛空裂<unclear reason="墨漬"><gap unit="char" extent="1"/></unclear>地皆振動</l>

The lower part of the character might be still visible so it allows you a guess. You can use <choice> again to record several conjectures and the cert attribute to assign probabilities to them.

 假使虛空裂
<choice>
<unclear cert="medium">天</unclear>
<unclear cert="high">大</unclear>
</choice>地皆振動

Here we are saying that we consider 大 at this position to be more likely than 天.

### 6. 4 Critical Apparatus

High quality editions of texts are indispensable for research in the Humanities. To create such editions editors use methods collectively called 经文校勘 (textual criticism). As is widely known, philology and textual criticism are important modes of thought in the Chinese tradition as well. In China during the apogee of 考證學 in the Qing dynasty Chinese scholars like 顧炎武 (1613 – 1682), 戴震 (1724-1777), and 段玉裁 (1735 – 1815) practiced incisive textual criticism on an unprecedented level. And just like their colleagues – the bible scholars and classicists in contemporary Europe – the proponents of 漢學 were led to a reappraisal of their own tradition. Textual scholarship therefore matters for the history of thought and a closer look at how authoritative editions are created in the digital medium is well worth the time.
Creating a new edition usually involves comparing previous 版本 (witnesses) of the text. Where the witnesses contain different (readings) the editor will in some ways record them in an (apparatus).
TEI offers mechanisms to allow editors to model existing editions and produce new ones.

The basic element for encoding the apparatus is <app> (apparatus) which contains the different <rdg> (readings) of a character or a passage. If you are working from a base text use <lem> (lemma) to record its readings. In a critical edition the editor not only records the variant readings, but also signals which reading is to be preferred on the basis of their scholarly acumen.
The following passage appears as 假使虛空裂天地皆振動 in an early 20th century edition of the Buddhist canon.

```
<l>假使虛空裂
<app>
        <lem wit="#【大】">天</lem>
        <rdg resp="#Taisho" wit="#【宋】#【聖】">大</rdg>
        <rdg resp="#CBETA" wit="#【麗】">大</rdg>
</app>
         地皆
<app n="fnT02p0382n09">
 <lem wit="#【大】">振</lem>
 <rdg resp="#Taisho" wit="#【宋】#【元】#【明】#【聖】">震</rdg>
</app>
                  動</l>
```

Here we say that the character 天 is used in the base text with the 符號 (siglum) "【大】". Other editions, however, designated as 【宋】【聖】have the reading 大 here. Similarly, the resp attribute says that "#Taisho" indicates a definition for the agency responsible for collating these witnesses. One more witness with the siglum【麗】was consulted by the agency called "#CBETA". Of course these sigla and abbreviations must be explained somewhere, typically in the TEI header (under <encodingDesc>).
The mechanisms sketched above is explained in more detail in Chapter 12 of the full guidelines. Next to support for critical editions the *linking* module will probably be needed for comparative editions (Ch 16). For editions that include syntactic or narrative analysis, an interesting topic which we do not have space to discuss here, see the Chapter on "Simple Analytic Mechanisms" (Ch.17) in the full guidelines.

### 6.5 Missing Characters

So-called 缺字(missing characters) have been a problem for CJK computer processing since its inception. The root of the problem is that, as in Europe and the US, governments and standard organizations in Asia were slow to react to the need to provide a common encoding standard for the representation of writing systems for computer processing and left it to the industry to devise encoding standards. As a result a multitude of mutually incompatible character encoding standards arose until the Unicode standard gradually emerged as the normative international standard. In the early 90s there were at least five different standard for Chinese, four for Korean and three for Japanese each of which contained a similar set of CJK characters. Since Unicode has only recently gained general acceptance in Asia there is a lot of legacy data to deal with. In Taiwan, for example, there is still a lot of legacy data in Big-5, the conversion of which to Unicode is not without problems. But compatibility problems aside, all legacy standards and even Unicode have difficulties solving the basic conundrum of CJK characters: there are so many of them. Unicode contains more than 70,000 CJK characters and will continue to add small numbers in the future. More than 40,000 of those are grouped in a section called Unicode Extension B, which was added to Unicode in 2001 (with version Unicode 3.1). But even with this large number of available characters, projects working with older printed material or inscriptions will have to deal with characters which are not in Unicode or need to annotate variants of existing

characters.

What now are 缺字 or, as they are called in Japanese, 外字? The expert definition is that these are characters missing in the character set. But in the experience of many users it means characters that are missing from the font they have chosen. With the advent of Unicode this difference in usage has become more visible. Even using Unicode many characters - although in theory part of the character set - are not generally available at the client end. Today there are only 3 or 4 fonts which can realize the 42,711 rare characters in Unicode CJK-Extension B. An online interface cannot assume the user has one of these fonts installed.[17] A project dealing with many rare characters will have to make decisions of how to solve certain issues on the encoding side as well as on the output side of the project.
On the encoding side, a project handling texts with rare characters will have to address the question of how to handle non-unicode characters and perhaps even how to annotate unicode characters that are difficult to process or need to be distinguished from similar variants.[18]
For instance the variants 說 (Unicode No. 63855) and 説 (Unicode No. 35500) can easily be encoded with the two distinct code-points provided by Unicode. For the character with the No.19756, however, there exist two variants:

and

Since Unicode makes no distinction between them and assigns the same code point to both, what appears on the screen depends on the operating system and the application used.

On the output or interface side, a project which aims at presenting its data, will have to display or otherwise identify non-Unicode and Unicode Extension B characters for the user, which in all likelihood will not have a font of sufficient size on her computer.
To a degree most of these problems can be solved by markup. TEI offers a mechanism to annotate "missing characters" within Unicode as well as non-Unicode characters.

To do so, TEI proposes a gaiji <g> element in the text. For example the verse line 既不得其味 [口*(隹/乃)]傷而虚還 contains a non-Unicode character:

既不得其味 □ 傷而虚還

In TEI code this might be expressed:

<l>既不得其味 <g ref="#id003"/>傷而虚還</l>[19]

The ref attribute refers to the description of the character in the TEI header, where information concerning the character is recorded in the <charDecl> contained in the <encodingDesc>. The

17 Another practical problem is that most operating systems and applications so far do not handle Unicode Extension B characters very well. Especially the Unicode solution of representing characters in UTF-16 with surrogate pairs is not well supported. Markup is not going to help us there, but projects handling rare Chinese characters should be aware of the issue.

18 Unicode encodes a large number of variants mainly for compatibility reasons, but also in order to respect historical and cultural developments in the writing of Chinese characters.

19 The <g> does not have to be empty, but can contain a preliminary mapping for easy display.

<charDecl> allows the encoders to add metadata for characters which, for whatever reason, merit additional information in the context of the project.

```
<encodingDesc>
…
<charDecl>
        <glyph xml:id="id003">[20]
                <glyphName>Non Unicode Character</glyphName>
                <charProp>
                        <localName>pronunciation</localName>
                        <value>zui3</value>
                </charProp>
                <mapping type="cbeta">[口*(隹/乃)]</mapping>
                <mapping type="cbetaNo">CB00047</mapping>
                <mapping type="regular">嘴</mapping>
                <graphic url="gaiji/g003.gif"/>
                <note>The character is listed in 高麗大藏經異體字字典, 漢城市：高麗大藏經研究所, 2000.
p. 132; #794. The pronunciation is given as zui3, the 正字 form as 嘴.</note>
        </glyph>
</charDecl>
</encodingDesc>
```

This allows information managers and interface designers a large number of possible mappings to draw on. Though the character is not contained in Unicode, using notes and mappings, the "missing character" is well documented in a standardized fashion, within the very document it appears. The <g> element does not have to be empty, depending on the use of the text some kind of mapping might be placed directly into the text.

### 6.6 Markup and Images

There are three main scenarios to relate TEI text to images:
First, in the context of a digitized work figures and images in the original can be embedded into the text similar to the <img> tag in HTML. The TEI element for this purpose is called <figure>, and allows for detailed information about the image to be included.
Secondly, in the context of aligning a full text in TEI with a digital facsimile, e.g. a scan or photograph, of the original. This <facsimile> mechanism, which links the text to the image is relatively new to TEI.

---

20 The TEI guidelines follow the Unicode standard in drawing a distinction between abstract, conceptual *characters* and realized, concrete *glyphs*. Accordingly, the Guidelines ask the user to decide whether a "missing character" is to be treated as a *character* or a variant *glyph* of an existing Unicode character. In the former case one is to use a <char> within the <charDecl>, in the latter <glyph>. Personally, I do not think that the character/glyph dichotomy is useful for conceptualizing how to handle CJK characters. 龍 竜 and 龙 are not glyphs of the same character in the way "*a*" and "a" are. Already Unicode has had to encoded innumerable "variants" (glyphs) both because of cultural-political as well as for technical reasons. What seemed a viable construct for theorizing about the units of a writing system – probably inspired by memories of platonic ideas (the glyph as the concrete manifestation of an abstract idea) – might work well for ABC and syllabic writing systems. In the context of CJK, however, is it often almost impossible to decide unambiguously whether the character in the text witness is a "true" *character* in the Unicode sense or a mere *glyph* variant of a existing character. Is the character in the example above a variant *glyph* of 嘴 or an "entirely distinct derived character (Ch.5.4)"? The adoption of the character/glyph distinction into TEI never seemed a good idea to me, but fortunately it matters little for actual encoding, whether you call the units in the <charDecl> <char>s or <glyph>s.

It is intended for use in the digitization of texts, such as manuscripts or inscriptions, where the material condition of the text make a facsimile desirable.
Thirdly, in recent years it has become possible to use TEI to annotate digital images. In this scenario, which we will not discuss in detail here, a digitized transcription is not needed or not possible (e.g. because the digital image is not an image of a text, but a painting). Nevertheless a project may want to associate annotations with an digital image in a consistent way, so that both the annotations and the images can be fully used in various processes. To align image files with TEI commentaries (usually via SVG) is a clever way to produce standardized annotations for image collections. In practice this usually involves a tool like the Image Markup Tool developed by Martin Holmes.[21]

1. Embedding information about an image in a TEI text.
To embed an image in a TEI document use the block element <figure>. To link to a image file include the empty element <graphic>. The target attribute on <graphic> has the same semantic value as the href attribute on <img> in HTML. <figure> can also contain <figureDesc>, a brief prose description of the image. This is similar to the ALT attribute on <img> in HTML. The <head> contained in <figure> is the place to record a caption that goes with the image.
To embed the TEI logo


we can write:

<figure>
<graphic target="/images/TEI-400.jpg"/>
<figDesc>Stylized yellow angle brackets with the letters<mentioned>TEI</mentioned> in between and <mentioned>text encoding initiative</mentioned> underneath.</figDesc>
<head>Figure 1: TEI Logo</head>
</figure>

This records that an image of the given description is in the text at this place, points to a image file, and records its caption.

2. Align a TEI text with a facsimile of the text
This mechanism was newly added in TEI P5 to accommodate the growing trend within the digitization of texts to offer the user a digital facsimile image next to the full text. The mapping of a TEI text to a digital facsimile is described in detail in Chapter 11 of the full guidelines (11. *Representation of Primary Structures*). There are two main mechanisms provided for this. The first is simply to put a facs attribute on a section of the text. Facs contains a URI to the image.

<pb n="p1" facs="scan001.png"/>

This provides interface-programmers with the basic pointers they need to align the full text with the image.

21 Available at http://tapor.uvic.ca/~mholmes/image_markup/.

The second way is somewhat more involved, but supports definition of distinct zones within a facsimile and ways of dealing with multiple images of the same page or inscription. This is especially needed in cases where the original is not a regular book page and the actual position of the text matters, as in the case of, for example, inscriptions or crossword puzzles. Say you would like to point out that the text is found at a certain location of the surface that carries the text. You would have to describe certain <zone>s on the facsimile image. This is done within the <surface> element of the top-level <facsimile>. For the details and some good examples see Chapter 11 of the full guidelines.
TEI allows only for the creation of rectangular zones; for polygons or rounded shapes the encoder can embed SVG information in the document. Roma provides a dedicated schema for TEI documents, which allows to embed SVG easily.

## 7. Become a member

TEI is a non-profit consortium that makes the standard freely available. It relies on membership fees to stay functional and maintain the standard in the future. If you use TEI and think it is a standard worth supporting you might consider joining individually or, even better, convince your institution to join the TEI consortium. Individuals join the consortium as subscribers. The benefits are no registration fees for the annual meeting and membership in a learned society, a much underrated pleasure these days. Institutional members gain voting rights for the TEI board and have considerable influence in the future development of the standard. Annual membership fees are comparatively low and vary according to size of institution and economic development bracket. So far the consortium has almost no members representing Asia, which is unfortunate considering Asia's rich textual heritage. Please consider joining the consortium. More information is found at: http://www.tei-c.org/Membership/join.xml.