Data Entry Specs 1.2

Version for Chinese Text

Wolfgang Schmidle, Martina Siebert, Martin Hofmann,
Klaus Thoden, Malcolm D. Hyman

Max Planck Institute for the History of Science, Berlin, Germany

11th December 2008

Contents

1 File Conventions

Save the text in plain text format (`.txt`) with Unicode `utf-8` encoding. If the text is saved in more than one file, number the parts, for example `Euclid_part_001.txt`, `Euclid_part_002.txt`, and so on. Create a `zip` archive of all files.

Make use of the complete character repertoire found in Unicode version 5.1.0.

This includes characters in the following Unicode blocks when applicable:

- CJK Unified Ideographs Extension A (U+3499 – U+4DFF),
- CJK Unified Ideographs Extension B (U+20000 – U+2A6DF),
- CJK Compatibility Ideographs Supplement (U+2F800 – U+2FA1F).

We will also need the list of unknown characters (see section 2.5.2). If the list is hand-written, scan it and save it as PDF file.

2 General Markup

Type the entire contents of one page, then go on to the next page. Do not mix the contents of different pages.

2.1 Page Breaks, Page Numbers and Running Heads

Page breaks are marked by `<pb>`. If the page has a page number, type it within the `<pb>` tag, e.g. `<pb 六>`. Type the page number exactly as it appears in the book. If there is a running head on the page, it is marked by `<rh>` and `</rh>`. Type the running head immediately after the `<pb>` tag.

Type the `<pb>` and `<rh>` tags before you type any page content.

The centre section of a traditional printing page (*banxin* 版心) is equivalent to a running head in a western book layout. In this case, repeat `<pb>` and the running head for each half-page, but add `a` and `b` to the page number, e.g. `<pb 三a>` and `<pb 三b>`, or `<pb a>` and `<pb b>` if there is no page number.

If the characters of the running head are cut off on the scanned page, type them anyway. Type large spaces in the running head as a single IDEOGRAPHIC SPACE character U+3000.

Please note: In the digitization of the book, the two half-pages may be on the same scan or on two consecutive scans.

Examples

<pb 三十二**a><rh>**泰西事物起原　第十一章**</rh>**

<pb 十二**a><rh>**事物攷辨卷之六十三　**<sm>**植物**</sm>**　帶經堂**</rh>**

<pb 一**a><rh>**閩產錄異　卷一**<sm>**穀屬**</sm></rh>**

<pb 一**b><rh>**閩產錄異　卷一**<sm>**穀屬**</sm></rh>**

⟶ for <sm> see section 3.2.1. An example of two complete half-pages with their running heads can be seen in section 2.2.2.

2.2 Text Blocks

Type a return after each line of the printed page.

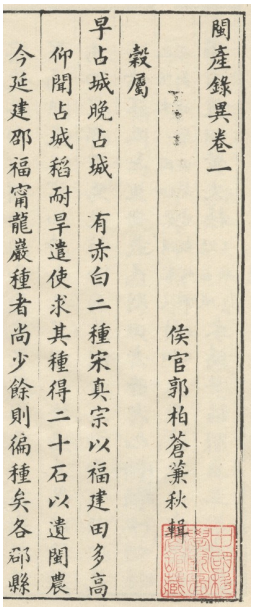
Do not insert a space at the end of the line.

2.2.1 Headings

Headings are marked by <h> and </h>. If you can identify a heading as the book title, use <ti> for the whole line. If you can identify a heading as the name of the author, compilator, proofreader etc. (*ti* 題), use <ti> too. If a text has different levels of headings, use <h 1>, <h 2>, and so on.

If a heading, book title, author, etc. is indented, do not mark this. Type large spaces in the heading as a single IDEOGRAPHIC SPACE character U+3000.

Example


閩產錄異卷一
侯官郭柏蒼蓀秋輯
穀屬
早占城晚占城　有赤白二種宋真宗以福建田多高
仰聞占城稻耐旱遣使求其種得二十石以遺閩農
今延建邵福甯龍巖種者尚少餘則徧種矣各郡縣

```
<ti>閩產錄異卷一</ti>
<stamp>
<ti>侯官郭柏蒼蓀秋輯</ti>
<h 1>穀屬</h>
<h 2>早占城晚占城</h>  <p>有赤白二種宋真宗以福建田多高
仰聞占城稻耐旱遣使求其種得二十石以遺閩農
今延建邵福甯龍巖種者尚少餘則徧種矣各郡縣
(some text)</p>
```

⟶ For <stamp> see section 2.4.2. For see section 2.2.2.

Please note: In this example, the base line for indentation (see section 2.2.2) would be one character below the printing frame.

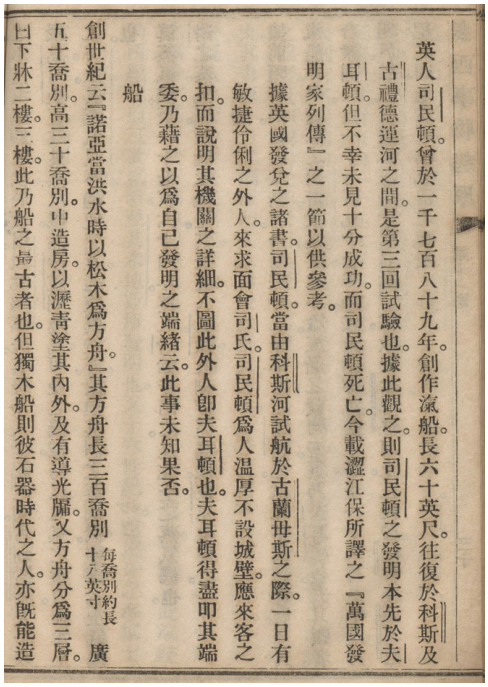
2.2.2 Paragraphs

Paragraphs are marked by and . Indented paragraphs are marked by one or more i, e.g. <p iii> for a paragraph that is indented by three character spaces. Outdented paragraphs are marked by one or more x, e.g. <p x>.

Every part of the book has a base line where all unindented paragraphs start. The base lines in the preface or table of contents may be different from the base line in the main text. Mark indentations relative to the base line. The indentation symbols i and x always refer to the first line of the paragraph. The remaining lines of the paragraph may have the same or a different indentation, which is not marked. If a paragraph is preceded by a sub-heading in the same line, as in the example in section 2.2.1, do not mark the indentation at all.

Make sure that for each there is a corresponding somewhere. If a paragraph starts and ends on different pages, the and tags are on these different pages.

Example

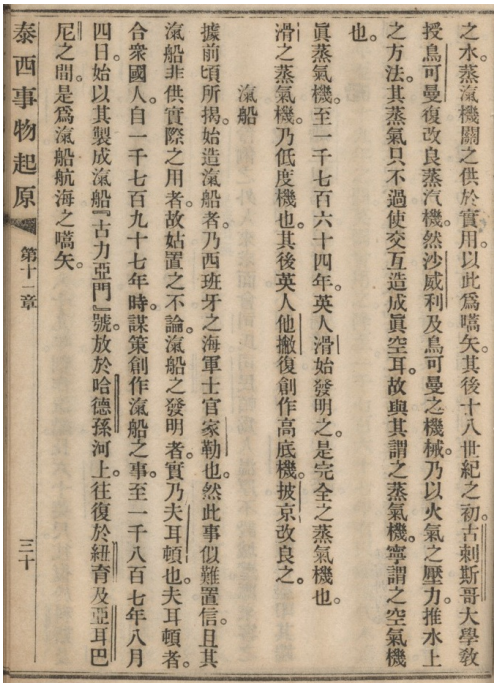


英人司民頓。曾於一千七百八十九年。創作瀛船。長六十英尺。往復於科斯及古禮德運河之間。是第三回試驗也。據此觀之。則司民頓之發明本先於夫耳頓。但不幸未見十分成功。而司民頓死亡。今載瀝江保所譯之『萬國發明家列傳』之一節以供參考。

據英國發兌之諸書。司民頓。當由科斯河試航於古蘭母斯之際。一日有敏捷伶俐之外人。來求面會司氏。司民頓為人溫厚不設城壁。應來客之扣。而說明其機關之詳細。不圖此外人卽夫耳頓也。夫耳頓得盡叩其端委。乃藉之以為自己發明之端緒云。此事未知果否。

船

創世紀云『諾亞當洪水時以松木為方舟』。其方舟長三百喬別（每喬別約長十八英寸）廣五十喬別。高三十喬別。中造房。以瀝青塗其內外。及有導光牖。又方舟分為三層。曰下牀二樓。三樓。此乃船之最古者也。但獨木船則彼石器時代之人。亦既能造



之水。蒸瀛機關之供於實用。以此爲嚆矢。其後十八世紀之初。古剌斯哥大學教授烏可曼復改良蒸汽機。然沙威利及烏可曼之機械。乃以火氣之壓力。推水上之方法。其蒸氣只不過使交互造成眞空耳。故與其謂之蒸氣機。寧謂之空氣機也。

眞蒸氣機。至一千七百六十四年。英人滑始發明之。是完全之蒸氣機也。滑之蒸氣機。乃低度機也。其後英人他撒復創作高底機。披京改良之。

瀛船

據前頃所揭。始造瀛船者。乃西班牙之海軍士官家勒也。然此事似難置信。且其瀛船非供實際之用者。故姑置之不論。瀛船之發明者。實乃夫耳頓也。夫耳頓者。合衆國人。自一千七百九十七年時。謀策創作瀛船之事。至一千八百七年八月四日。始以其製成瀛船『古力亞門』號。放於哈德孫河上。往復於紐育及亞耳巴尼之間。是爲瀛船航海之嚆矢。

泰西事物起原　第十一章　三十

<pb 三十a><rh>泰西事物起原 <sm>第十一章</sm></rh>
之水。蒸瀛機關之供於實用。以此為嚆矢。其後十八世紀之初。**<dl>**古剌斯哥**</dl>**大學教
授**<sl>**烏可曼**</sl>**復改良蒸汽機。然**<sl>**沙威利**</sl>**及烏可曼之機械。乃以火氣之壓力。推水上
之方法。其蒸氣只不過使交互造成眞空耳。故與其謂之蒸氣機。寧謂之空氣機
也。** **
** **眞蒸氣機。至一千七百六十四年。英人**<sl>**滑**</sl>**始發明之。是完全之蒸氣機也。
<sl>滑**</sl>**之蒸氣機。乃低度機也。其後英人**<sl>**他撒**</sl>**復創作高底機。**<sl>**披京**</sl>**改良之。** **
<h 2>瀛船**</h>**
** **據前頃所揭。始造瀛船者。乃西班牙之海軍士官**<sl>**家勒**</sl>**也。然此事似難置信。且其
瀛船非供實際之用者。故姑置之不論。瀛船之發明者。實乃**<sl>**夫耳頓**</sl>**也。夫耳頓者。
合衆國人。自一千七百九十七年時。謀策創作瀛船之事。至一千八百七年八月
四日。始以其製成瀛船『古力亞門』號。放於**<dl>**哈德孫**</dl>**河上。往復於**<dl>**紐育**</dl>**及**<dl>**亞耳巴
尼**</dl>**之間。是為瀛船航海之嚆矢。** **

<pb 三十b><rh>泰西事物起原 <sm>第十一章</sm></rh>
<p i>英人**<sl>**司民**</sl>**頓。曾於一千七百八十九年。創作瀛船。長六十英尺。往復于**<dl>**科斯**</dl>**及
<dl>古禮**</dl>**德運河之間。是第三回試驗也。據此觀之。則**<sl>**司民頓**</sl>**之發明本先於**<sl>**夫
耳**</sl>**頓。但不幸未見十分成功。而司民頓死亡。今載瀝江保所譯之『萬國發
明家列傳』之一節以供參考。** **
<p iii>據英國發兌之諸書。**<sl>**司民**</sl>**頓。當由**<dl>**科斯**</dl>**河試航於**<sl>**古蘭母斯**</sl>**之際。一日有
敏捷伶俐之外人。來求面會**<sl>**司**</sl>**氏。**<sl>**司民頓**</sl>**為人溫厚不設城壁。應來客之
扣。而說明其機關之詳細。不圖此外人卽夫**<sl>**耳頓**</sl>**也。夫耳頓得盡叩其端
委。乃藉之以為自己發明之端緒云。此事未知果否。** **
<h>船**</h>**
** **創世紀云『諾亞當洪水時以松木為方舟』。其方舟長三百喬別**<sm>**每喬別約長****十八英寸**<sm>** 廣
五十喬別。高三十喬別。中造房。以瀝青塗其內外。及有導光牖。又方舟分為三層。
曰下牀二樓。三樓。此乃船之最古者也。但獨木船則彼石器時代之人。亦 能造

⟶ For <sm> see section 3.2.1. For <sl> and <dl> see section 3.2.2.

2.3 Structured Text

2.3.1 Tables

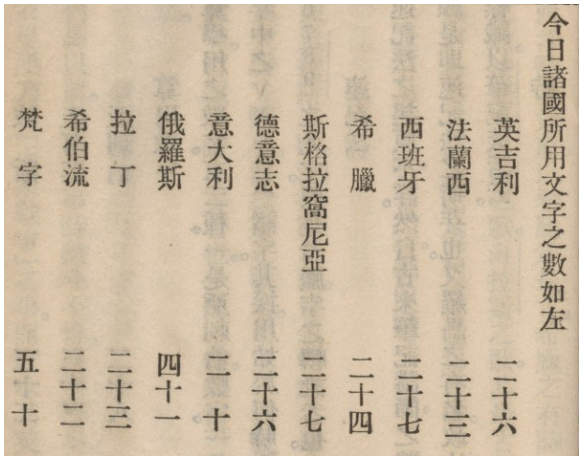
A table is marked by <tb> and </tb>. Use # as field separators. Type a return after each row. Do not type horizontal or vertical lines.

Type the <tb> and </tb> tags on separate lines. Do not mark indentations. The field separators may be lines or large spaces.

If you can identify a single space within a name etc. as a decorative space to make the table layout optically more pleasing, do not type it.

Example

```
<p>今日諸國所用文字之數如左</p>
<tb>
英吉利 # 二十六
法蘭西 # 二十三
西班牙 # 二十七
希臘 # 二十四
斯格拉窩尼亞 # 二十七
德意志 # 二十六
意大利 # 二十
俄羅斯 # 四十一
拉丁 # 二十三
希伯流 # 二十二
梵字 # 五 十
</tb>
```


今日諸國所用文字之數如左
英吉利 二十六
法蘭西 二十三
西班牙 二十七
希臘 二十四
斯格拉窩尼亞 二十七
德意志 二十六
意大利 二十
俄羅斯 四十一
拉丁 二十三
希伯流 二十二
梵字 五十

2.3.2 Lists

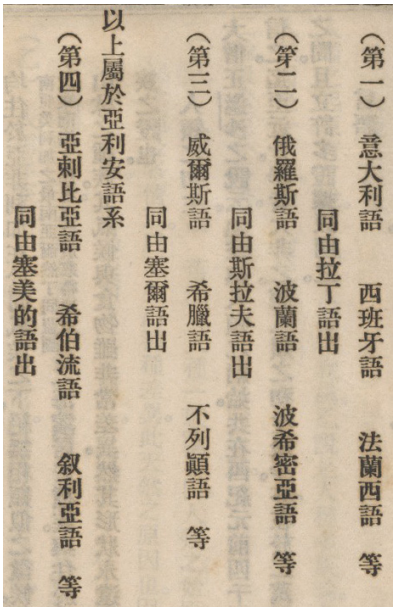
A list is marked by <list> and </list>. Use # for large spaces, if there are any.

Type the <list> and </list> tags on separate lines. If the items on consecutive text lines belong to the same list entry, use # at the beginning of the next line. Do not mark indentations.

Unlike in tables, type each single space, i.e. do not omit single spaces even if they seem to be merely decorative.

Example

```
<list>
（第一）　意大利語 # 西班牙語 # 法蘭西語　等
# 同由拉丁語出
（第二）　俄羅斯語 # 波蘭語 # 波希密亞語　等
# 同由斯拉夫語出
（第三）　威爾斯語 # 希臘語 # 不列顚語　等
# 同由塞爾語出
</list>
<p>以上屬於亞利安語系</p>
<list>
（第四）　亞剌比亞語 # 希波流語 # 叙利亞語　等
# 同由塞美的語出
</list>
```


（第一）意大利語　西班牙語　法蘭西語　等
同由拉丁語出
（第二）俄羅斯語　波蘭語　波希密亞語　等
同由斯拉夫語出
（第三）威爾斯語　希臘語　不列顚語　等
同由塞爾語出
以上屬於亞利安語系
（第四）亞剌比亞語　希伯流語　叙利亞語　等
同由塞美的語出

⟶ For an example of a list-like structure without large spaces see section 2.3.3 (second example).

2.3.3 Tables of Contents

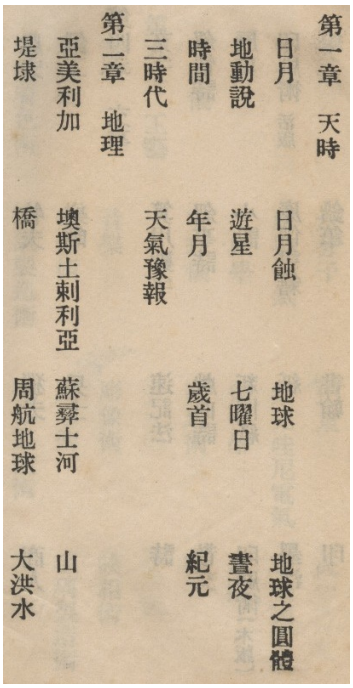
A table of contents is marked by <toc> and </toc>. If the table of contents has a table-like structure or a list-like structure with large spaces, use #.

Type the <toc> and </toc> tags on separate lines. A table of contents may look like a table or like a list.

Examples

a table-like table of contents

a list-like table of contents without large spaces




```
<toc>
<h>第一章　天時</h>
日月 # 日月蝕 # 地球 # 地球之圓體
地動說 # 遊星 # 七曜日 # 晝夜
時間 # 年月 # 歲首 # 紀元
三時代 # 天氣豫報
<h>第二章　地理</h>
亞美利加 # 墺斯土剌利亞 # 蘇彝士河 # 山
堤埭 # 橋 # 周航地球 # 大洪水
(some text)
</toc>
```

```
<toc>
(some text)
安南稻　米麥　牛尾粟<sm>雀粟　鵞掌粟　狗尾\\粟　虎尾粟　黃粟</sm>
黃粱　釣鉤黍<sm>馬尾黍　番黍　鴨脚黍\\　黑黍　長芒黍　膏黍</sm>　豆<sm>白豆\\　黃豆　黑豆　綠豆　豇豆　豌豆　赤小豆　青豆\\　褐豆　刀豆　虎爪豆　蠍眼豆　皂莢豆　羊
(some text)</sm>
</toc>
```

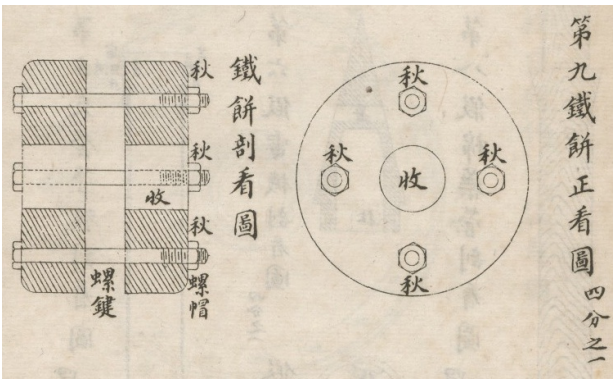
2.4 Printed Images

2.4.1 Figures

Where a figure occurs in the text, type a `<fig>` tag on a separate line. If you can identify a caption of the figure, mark it by `<cap> </cap>`. Additional text that describes parts of the figure is marked by `<desc> </desc>`.

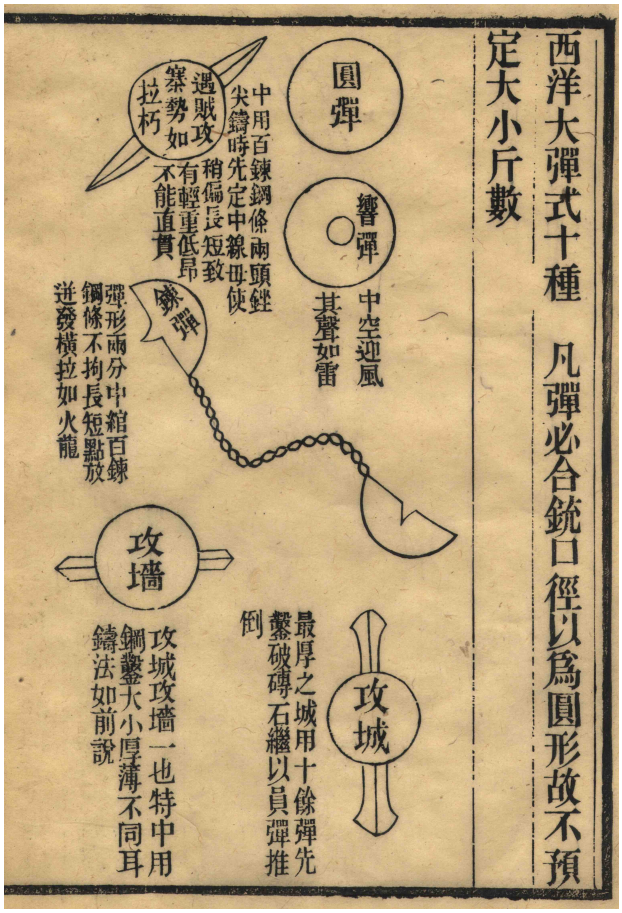
Type the caption on a separate line after `<fig>`. A figure may have more than one description. Type each description on a separate line after `<fig>` and `<cap>`. If the same description is repeated in a figure, type it only once.

Examples



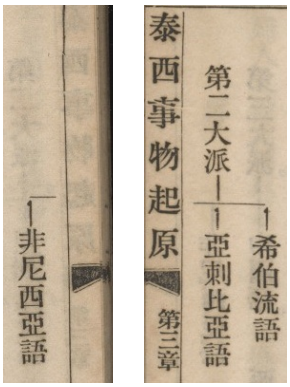

```
<fig>
<cap>第九鐵餅正看圖<sm>四分之一</sm></cap>
<desc>秋</desc>
<desc>收</desc>
<fig>
<cap>鐵餅剖看圖</cap>
<desc>秋</desc>
<desc>螺帽</desc>
<desc>收</desc>
<desc>螺鍵</desc>
```

西洋大彈式十種　凡彈必合銃口徑以爲圓形故不預
定大小斤數
<fig>
<desc>中空迎風
其聲如雷</desc>
<desc>圓彈</desc>
<desc>響彈</desc>
<fig>
<desc>中用百鍊鋼條兩頭銼
尖鑄時先定中線毋使
稍偏長短致
有輕重低昂
不能直貫</desc>
<desc>遇賊攻
寨勢如
拉朽</desc>
<fig>
<desc>彈形兩分中綰百鍊
鋼條不拘長短點放
迸發橫拉如火龍</desc>
<desc>鍊彈</desc>
<fig>
<desc>最厚之城用十餘彈先
鑿破磚石繼以員彈推
倒</desc>
<desc>攻城</desc>
<fig>
<desc>攻城攻墻一也特中用
鋼鑿大小厚薄不同耳
鑄法如前說</desc>
<desc>攻墻</desc>



<pb a><rh>泰西事物起原　<sm>第三章</sm></rh>
(some text)
<fig>
<desc>第二大派</desc>
<desc>希伯流語</desc>
<desc>亞剌比亞語</desc>

<pb b><rh>泰西事物起原　<sm>第三章</sm></rh>
<desc>非尼西亞語</desc>
(some text)



2.4.2 Stamps

Stamps are marked by `<stamp>`. Type the `<stamp>` tag on a separate line. Do not type the text in the stamp.

⟶ For an example see section 2.2.1.

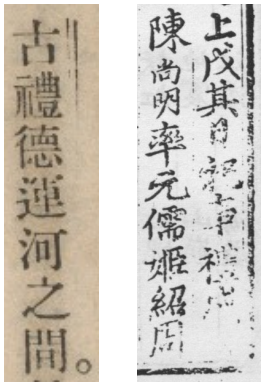
2.5 Unreadable Text

2.5.1 Characters You are Unsure About

If you are not sure about a character, type <?> after it. If you are unsure about a whole paragraph, type <?> directly after the <p> tag, i.e. <p><?>. A completely unreadable character is typed as @. If many characters are unreadable, use <gap> instead of @.

Use one @ for each unreadable character, e.g. unr@@dable. If in doubt, use <gap>, e.g. unr<gap>dable. If you are unsure about a group of characters, for example a whole word, do not type <?> repeatedly for every character, e.g. type word<?> rather than w<?>o<?>r<?>d<?>.

Examples



<dl>古禮**</dl><?>**德運河之間。

上戊其日祀事禮成**<?>@**
陳尚明率元儒姬紹周**<?>**

Please note: In the first example, the characters are readable but the double line (see section 3.2.2) is badly printed.

⟶ For unknown rather than unreadable characters please refer to section 2.5.2.

2.5.2 Unknown Characters

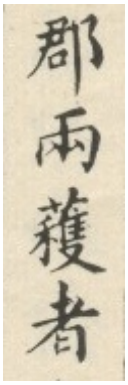
If there is an unknown character in the text, or if the character is not included in Unicode 5.1.0, add it to the numbered list of unknown characters. From then on, type its number whenever it occurs in the text, e.g. <001>.

Before you create a number for an unknown character, first check whether it is already on the list of unknown characters. Assign the number <001> to the first unknown character, <002> to the second unknown character, and so on. Do not assign the same number twice. Use this number to type the unknown character. Always use the same number if the same unknown character occurs again.

Please note: Make sure that for a given text there is a single list containing all unknown characters, and that everyone uses this list. When the text is sent back to us, we will need a copy of this list. (See also section 1.)

Example

a character not included in Unicode 5.1.0



郡兩**<001>**者

⟶ For unreadable rather than unknown characters please refer to section 2.5.1.

3 Chinese Characters

3.1 General

3.1.1 Punctuation

Type the punctuation to the right of characters.

⟶ For an example see section 2.2.2.

3.1.2 Spaces

Type spaces in Chinese text as the IDEOGRAPHIC SPACE character U+3000.

In running heads (`<rh>`, see section 2.1) and headings (`<h>` and `<ti>`, see section 2.2.1), type large spaces as a single ideographic space character. In tables and lists (`<tb>`, `<list>` and `<toc>`, see section 2.3), use `#` for large spaces.

If you encounter a large space in a normal paragraph (`<p>`, see section 2.2.2), make sure that none of the cases above apply. If it is indeed a normal paragraph, type the large space as more than one ideographic space, according to its length.

⟶ For an example of large spaces that are typed as a single ideographic space character, see section 2.1. For an example of large spaces that are typed as `#`, see section 2.3.1.

3.2 Type Styles

3.2.1 Small Characters

Strings of small characters are marked by `<sm> </sm>`. Indicate half-column breaks by `\\`.

⟶ For an example see section 2.3.3 (second example). This examples includes strings of small characters over more than one line of text.

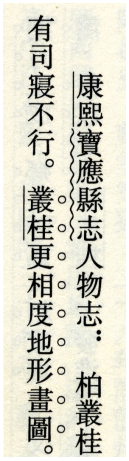
3.2.2 Underlinings

A single line next to characters is marked by <sl> </sl>. A double line next to characters is marked by <dl> </dl>. A circled line next to characters is marked by <cl> </cl>. A wavy line next to characters is marked by <wl> </wl>.

Please note: In old texts, the lines are to the right of characters. In modern texts, the lines may also be to the left of characters. The position to the left or right is not encoded.

Example

underlinings to the left and right



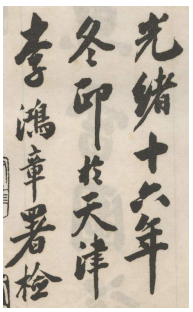
<sl>康熙**</sl><wl>**寶應縣志**</wl>**人物志：　柏叢桂
有司寢不行。**<cl><sl>**叢桂**</sl>**更相度地形畫圖**</cl>**。

⟶ For an example with underlinings to the right of characters see section 2.2.2. This example includes <dl> for double lines.

3.2.3 Individualized Character Style

A paragraph in an individualized character style is marked by ics in the tag, i.e. <p ics>.

Example



<p ics>光緒十六年
冬印於天津
李鴻章署检** **