Twelve gazetteers from the archive have appeared in a print series, which in the preface is described as "a snapshot … in the development of the digital text". (Bingenheimer 馬德偉 2013: Vol.1: vii). In the present paper we want to highlight another use of the corpus: as benchmark for research on NER in classical Chinese.

As of today (Feb. 2015) the corpus consists of two sections: The texts of 208 Temple Gazetteers are available without punctuation and only basic structural TEI markup. Besides those, another 15 gazetteers were digitized with added punctuation and sophisticated, manual markup that identifies and disambiguates names and dates and links them to an authority database.

The 208 gazetteers are basically text-only versions with a metadata header and markup indicating distinct texts ( ) within the gazetteer, their headings (<heading>), authors (<byline>), page- and line-breaks (<pb>, <lb>). The markup differentiates prose (<p>) from table content (<table> a .o.), and verse sections (<lg> a. o.). Furthermore it aligns the page-breaks with the image files that are part of the archive distributables. Metadata on the image files is provided in the METS meta-data wrapper, which too is part of the archive. The combined character count of the 208 gazetteers with basic markup is c. 14,000,000[l].

The 15 gazetteers with modern punctuation are built on the same basic tag set, but include additional markup going far beyond it. The research team has painstakingly identified all person and location names (<persName>, <placeName>) as well as all dates (<date>) in the texts. KEY attributes link the names to authority databases (http://authority.ddbc.edu.tw/) that were created at Dharma Drum in the process of this and other projects. All name and date entities are not only identified, but also fully disambiguated. The name "Avalokiteśvara", for instance, might appear in the sources transcribed or translated in various ways: 阿縛盧枳低濕伐羅 *Afuluzhidishifaluo*, 觀音 *Guanyin*, 觀世音 *Guanshiyin*, 光世音 *Guangshiyin*, 觀自在 *Guanzizai*. These names are all mapped to the same ID (A002803), as are other designations of this Bodhisattva, such as 白衣大士 *Baiyi dashi* "white-robed Mahāsattva/Great Being" or 普陀大士 *Putuo dashi* "Mahāsattva/Great Being of Mount Putuo".

The process is identical to disambiguation in European languages. The pre-standardized spelling of Shakespeare – Shakespear, Shakspeare, Shakspere a.o.– all refer to the same person, as does "The Bard of Avon". Beyond merely recognizing NEs as part of POS tagging, which was the original problem of NER, the successful disambiguation of NEs at one point has to include some form of authority data or ontology that can assist the algorithm or, as in our case, the human encoder.

This separation between markup and authority database should be considered best practice as it reduces complexity. Certain distinctions that in NLP are usually implemented in the tagging itself, can be outsourced to the authority databases, for instance whether a person is considered historical or fictional, or a location is a city or a building (cf. Xiong et al. 2013).

The combined character count of the 15 fully marked-up gazetteers is c. 1,600,000[m]. For the amount of NEs and the ratio of NEs to the whole text see Tables 1 and 4 below. Within those 15 gazetteers, rare characters that appear in the woodblock prints were marked-up in TEI in a way that allows regularizing them in different output modes. For the print version, for instance, we were able to map rare variants to their closest

Table 1 Person and Location Entities

Gazetteer (Date)	Person-names (total)	Person-names (unique values)	Person-frequency factor (All names/Unique values)	Location-names (total)	Location-names (unique values)	Location- frequency factor ratio
重修普陀山志 *Chongxiu putuo shan zhi* "Revised Gazetteer of Mount Putuo" (1607)	1164	464	2.51	1584	414	3.83
明州阿育王山志 *Mingzhou ayuwang shan zhi* "Mingzhou Aśoka Temple Gazetteer" (1612)	2741	679	4.04	1956	498	3.93
福建泉州開元寺志 *Fujian quanzhou kaiyuan si zhi* "Fujian Quanzhou Kaiyuan Temple Gazetteer" (1643)	767	320	2.40	548	230	2.38
明州阿育王山續志 *Mingzhou ayuwang shan xuzhi* "Supplement to the Mingzhou Aśoka Temple Gazetteer" (1757)	1785	526	3.39	1104	285	3.87
黃檗山志 *Huangbo shan zhi* "Mount Huangbo Gazetteer" (1824)	1988	582	3.42	1784	385	4.63
慧因高麗華嚴教寺志 *Huiyin gaoli huayanjiao si zhi* "Huiyin Koryo Huanyan School Temple Gazetteer" (1881)	1422	429	3.31	1080	279	3.87
天台山方外志 *Tiantai shan fangwai zhi* "Mount Tiantai Gazetteer" (1894)	6724	1768	3.80	6736	1385	4.86
寒山寺志 *Hanshan si zhi* "Hanshan Temple Gazetteer" (1922)	1751	476	3.68	1022	281	3.64
普陀洛迦新志 *Putuoluojia xin zhi* "New Gazetteer of Mount Putuo" (1924)	5856	2272	2.58	5452	1276	4.27
清涼山志 *Qingliang shan zhi* "Mount Qingliang Gazetteer" (1933)	2981	663	4.50	2634	613	4.30
峨眉山志 *Emei shan zhi* "Mount Emei Gazetteer" (1934)	2696	808	3.34	5078	929	5.47
九華山志 *Jiuhua shan zhi* "Mount Jiuhua Gazetteer" (1938)	3025	916	3.30	4300	821	5.24

equivalent in Unicode, thus avoiding having to font a large number of variants (see Bingenheimer 馬德偉 2013: 凡例 *fanli* "Editorial principles").

The project was constructed as digital archive, and the same code is used in different output formats. First an example of the XML/TEI master format which in a text editor might appear like this (see Figure 1). **F1**

The KEY attributes on the <persName>, <placeName> and <date> elements connect the named entity to the authority databases. The XML/TEI data is transformed to HTML for an online interface that can exploit these connections as links, which allow users to access the information from the database (see Figure 2). **F2**

In the printed edition on-click links are difficult to implement, instead person and place indices are compiled and appended to each volume. Dates are generally mapped to the (proleptic) Gregorian calendar (see Figure 3). **F3**

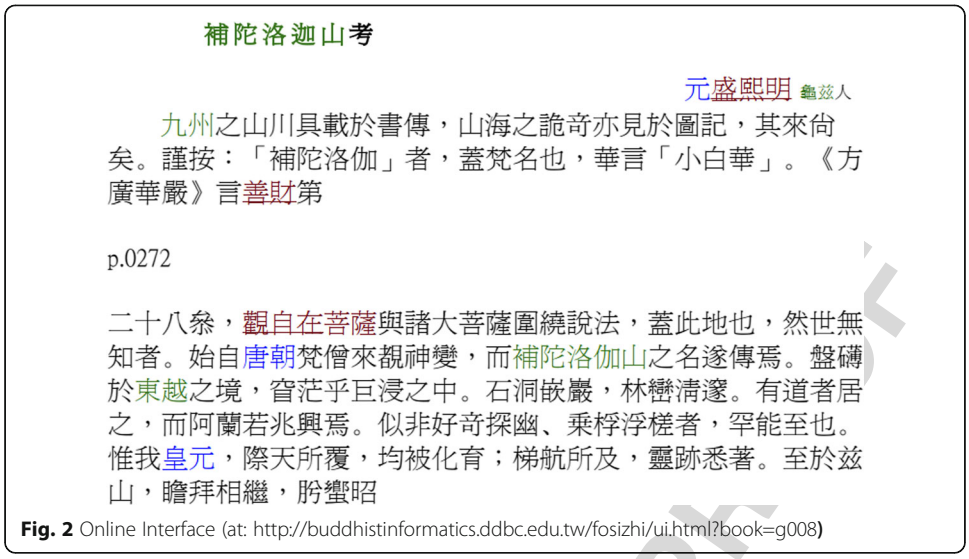
5 Named Entities and NER in the Digital Archive of Buddhist Temple Gazetteers

The separation of semantic information (encoded in TEI) from the presentation layer is the basis for creating different views of the texts (print, online, as audio book etc.). However, markup also allows a more sophisticated, research-oriented analysis of the texts. We can now answer questions such as: How much of the text consists of NEs? How many NEs are there on average per text? Are the averages comparable, or do they vary significantly between texts? How many unique persons and places are there as compared with the number of overall occurrences? Below are the query results and some answers to these questions[n] (see Table 1).

Table 1 is about occurrences of person and location names in the gazetteers. The markup allows a similar analysis for date expressions, opening up the prospect of visualizing events along timelines, but for now we will limit ourselves to person and place names. The KEY attribute on <persName> and <placeName> that points to unique authority database entries, allows to remove multiple mentions as well as homonymy. Counting unique values results in the actual number of persons and locations mentioned in the text with great precision (there are a few "unknown" names which resist identification). The range that is described by the "Frequency factor" approximates how many times person or place names are mentioned in the text on average. The higher the ratio the more often a gazetteer repeats its names. The ratio varies slightly between

```
<div>
<head><placeName key="PL000000013847">補陀洛迦山</placeName>考</head>
<byline><date key="j21856422220977" notBefore-iso="1271-12-25" notAfter-iso=
"1368-09-22">元</date><persName key="A001122">盛熙明</persName><seg rend=
"font-size:small"><placeName key="PL000000047814">龜茲</placeName>人</seg></byline>
<p><placeName key="PL000000000083">九州</placeName>
之山川具載於書傳，山海之詭奇亦見於圖記，其來尚矣。謹按：「補陀洛伽」者，蓋梵名也，華言「小白華」。《方廣華嚴
》言<persName key="A007249">善財</persName>第<pb facs="1B009P291.jpg" n="0272"/>二十八叅，
<persName key="A002803">觀自在<roleName>菩薩</roleName></persName>
與諸大菩薩圍繞說法，蓋此地也，然世無知者。始自<date key="j19469512052491" notBefore-iso=
"0618-06-21" notAfter-iso="0907-06-06">唐朝</date>梵僧來覩神變，而<placeName key=
"PL000000013847">補陀洛伽山</placeName>之名遂傳焉。盤礴於<placeName key="PL000000008353">東越
</placeName>
之境，窅茫乎巨浸之中。石洞嵌巖，林巒清邃。有道者居之，而阿蘭若兆興焉。似非好奇探幽、乘桴浮槎者，罕能至也。惟
我<date key="j21856422220977" notBefore-iso="1271-12-25" notAfter-iso="1368-09-22">皇元
</date>，際天所覆，均被化育；梯航所及，靈跡悉著。至於茲山，瞻拜相繼，肹蠁昭
```

Fig. 1 Code (XML/TEI markup. Data available at: http://buddhistinformatics.ddbc.edu.tw/fosizhi/) f1.1 Q8


補陀洛迦山考

元盛熙明 龜茲人

九州之山川具載於書傳，山海之詭奇亦見於圖記，其來尚矣。謹按：「補陀洛伽」者，蓋梵名也，華言「小白華」。《方廣華嚴》言善財第

p.0272

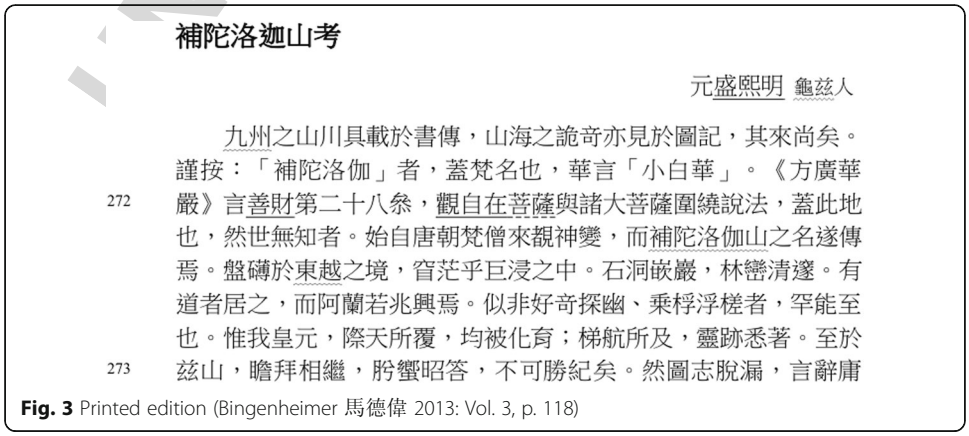
二十八叅，觀自在菩薩與諸大菩薩圍繞說法，蓋此地也，然世無知者。始自唐朝梵僧來覩神變，而補陀洛伽山之名遂傳焉。盤礴於東越之境，窅茫乎巨浸之中。石洞嵌巖，林巒清邃。有道者居之，而阿蘭若兆興焉。似非好奇探幽、乘桴浮槎者，罕能至也。惟我皇元，際天所覆，均被化育；梯航所及，靈跡悉著。至於茲山，瞻拜相繼，肹蠁昭

Fig. 2 Online Interface (at: http://buddhistinformatics.ddbc.edu.tw/fosizhi/ui.html?book=g008) f2.1 Q9

gazetteers: 2.4 to 4.5 with person names; and 2.38 to 5.47 with location names. The lower limits are probably due to the shortness of the 福建泉州開元寺志. The arithmetic average might not be the most useful average measure here. All gazetteers have a "long tail" of many unique occurrences of names, and a small number of names that are mentioned very frequently. More sophisticated math will be able to describe the distribution more precisely.

In a next step, by using the (freely available) Dharma Drum person authority database (http://authority.ddbc.edu.tw/person/) further queries can answer which persons were Q11 mentioned most frequently. Below are the twenty most frequent person names in the early 20th century gazetteers of the three most famous Buddhist Mountains: Mount Wutai, Mount Emei and Mount Putuo[o] (see Table 2). **T2**

Studying these lists reveals who the gazetteer compiler deemed important for the site. Among the most frequent names are Bodhisattvas, famous abbots, monks and laypersons associated with the site. An immediate, if trivial, result is that one can easily spot which Bodhisattva is associated with which site. On a second glance, somewhat less obvious, it appears that the three sites differ in the prominence they give to *other* Bodhisattvas and Buddhas. Thus in the 峨眉山志, besides 普賢菩薩 *Puxian pusa*


補陀洛迦山考

元盛熙明 龜茲人

九州之山川具載於書傳，山海之詭奇亦見於圖記，其來尚矣。謹按：「補陀洛伽」者，蓋梵名也，華言「小白華」。《方廣華
272 嚴》言善財第二十八叅，觀自在菩薩與諸大菩薩圍繞說法，蓋此地也，然世無知者。始自唐朝梵僧來覩神變，而補陀洛伽山之名遂傳焉。盤礴於東越之境，窅茫乎巨浸之中。石洞嵌巖，林巒清邃。有道者居之，而阿蘭若兆興焉。似非好奇探幽、乘桴浮槎者，罕能至也。惟我皇元，際天所覆，均被化育；梯航所及，靈跡悉著。至於
273 茲山，瞻拜相繼，肹蠁昭答，不可勝紀矣。然圖志脫漏，言辭庸

Fig. 3 Printed edition (Bingenheimer 馬德偉 2013: Vol. 3, p. 118) f3.1

"Samantabhadra", there is also 釋迦牟尼佛 *Shijiamoni fo* "Shakyamuni", 文殊菩薩 *Wenshu pusa* "Mañjuśrī", 觀世音菩薩 *Guanshiyin pusa* "Avalokiteśvara", and 普眼菩薩 *Puyan pusa* "Samantanetra", among the twenty most frequently mentioned names. The 清涼山志, besides Mañjuśrī, frequently mentions 普賢菩薩 *Puxian pusa* "Samantabhadra" and 阿彌陀佛 *Amito fo* "Amitabha Buddha". In contrast, the list for the 普陀洛迦新志 includes no other savior figures apart from Avalokiteśvara. It appears that Mount Putuo is more exclusively associated with Avalokiteśvara than the other two sites are with Mañjuśrī and Samantabhadra respectively.

That Mount Putuo has been more focused on Avalokiteśvara than other Buddhist mountains on "their" Bodhisattva, is corroborated by a look at the temple architecture and other features of the site. Guanyin is the central image in the main hall of the major temple at Mount Putuo and the other two major temples too are tied to the Guanyin cult. Both Mount Wutai and Mount Emei accommodate a larger and more diverse number of temples, and are iconographically less committed to a single savior figure.

Another question that can be asked is regarding the connection of the sites with literary figures. In the case of Mount Wutai the name of 張商英 *Zhang Shangying* "Zhang Shangying" is firmly associated with the site, because of his travelogue (Gimello 1992). In the 峨眉山志, 蘇軾 *Su Shi* "Su Shi" appears prominently, but not 范成大 *Fan Chengda* "Fan Chengda", who wrote an important travelogue (Fan appears only on position 31) (Hargett 2006). Among the "Top 20" of the Putuo Gazetteer, on the other hand, there is no famous literary figure.

In cases where a site has gazetteers of different periods we can attempt a diachronic view and compare how the "landscape of names" has changed over the centuries (see Table 3). **T3**

Apart from legendary and religious figures (觀音 *Guanyin* "Avalokiteśvara", 善財 *Shancai* "Sudhana", 梅福 *Mei Fu* "Mei Fu") only two names have made it into the list both in the early 17th and the 20th century gazetteer. Although neither 真融 *zhenrong* "Zhenrong" (1524–1592) nor 真歇清了 *Zhenxie Qingliao* "Zhenxie Qingliao" (1088–1151) are household names, the above highlights them as important for the cultural memory of Mount Putuo and might guide further research in their direction.

Another question it now becomes possible to ask is: How much of the text consists of named entities and dates? This is important for comparative studies between corpora or to flag single gazetteers that have "eccentric" NE patterns. For our twelve gazetteers the tally is shown in Table 4. **T4**

The result shows that temple gazetteers in the last three centuries have a similar text/name-date ratio. The overall range of 9.9 % to 14.3 % can be narrowed significantly by disregarding two outliers, the 黃檗山志 (g086) and the 明州阿育王山續志 (g011). The 黃檗山志 is focused on the figure of Yinyuan and his disciples and therefore mentions fewer different names, using pronouns and demonstratives, which lowers the character count for proper names. The 明州阿育王山續志 consists almost entirely of poetry. Most of the short poems are associated with a place and an author, which might account for the relative high amount of names and dates. Without these two the range is between 10.3 % and 13.3 %, merely 3 percent. Is this typical for multi-genre compilations of classical Chinese texts? We do not know, because the gazetteer corpus is so far the only freely accessible corpus of classical Chinese that is tagged for NEs. How

Table 2 Twenty most frequent person names in three gazetteers t2.1

普陀洛迦新志(1924)	清涼山志(1933)	峨眉山志 (1934)	t2.2
Person-names occurances	Person-names occurances	Person-names occurances	t2.3
觀世音菩薩 *Guanshiyi pusa* (x 167)	文殊菩薩 *Wenshu pusa* (x 365)	普賢菩薩 *Puxian pusa* (x 308)	t2.4
		(unknown x 87)	t2.5
性統 *Xingtong* (x 118)	(unknown x 189)	文殊菩薩 *Wenshu pusa* (x 43)	t2.6
通旭 *Tongxu* (x 99)	張商英 *Zhang Shangyin* (x 68)	福登 *Fudeng* (x 41)	t2.7
繹堂 *Yitang* (x 73)	鎮澄 *Zhencheng* (x 63)	克勤 *Keqin* (x 37)	t2.8
法澤 *Faze* (x 53)	德清 *Deqing* (x 51)	廣成子 *Guang Chengzi* (x 33)	t2.9
化聞 *Huawen* (x 49)	真可 *Zhenke* (x 48)	可聞 *Kewen* (x 29)	t2.10
(unknown x 47)[a]	福登 *Fudeng* (x 44)	孫思邈 *Sun Simiao* (x 27)	t2.11
藍理 *Lanli* (x 43)	無著 *Wuzhuo* (x 38)	蔣超 *Jiang Chao* (x 25)	t2.12
梅福 *Mei Fu* (x 41)	法照 *Fazhao* (x 33)	黃帝 *Huangdi* (x 24)	t2.13
真融 *Zhenrong* (x 38)	袾宏 *Zhuhong* (x 28)	胡世安 *Hu Shian* (x 23)	t2.14
愛新覺羅玄燁 *Aixingjueluo Xuanye* (x 38)	普賢菩薩 *Puxian pusa* (x 26)	釋迦牟尼佛 *Shijiamuni fo* (x 23)	t2.15
裘璉 *Qiu Lian* (x 37)	澄觀 *Chengguan* (x 23)	蘇軾 *Su Shi* (x 20)	t2.16
印光大師 *Yinguang dashi* (x 35)	義存 *Yizun* (x 22)	通天 *Tongtian* (x 20)	t2.17
善財童子 *Shancai tongzi* (x 35)	道義 *Daoyi* (x 22)	樵陽子 *Qiao Yangzi* (x 20)	t2.18
開如 *Kairu* (x 35)	阿彌陀佛 *Amituo fo* (x 21)	呂洞賓 *Lü Dongbin* (x 19)	t2.19
了餘 *Liaoyu* (x 35)	法本 *Faben* (x 20)	慧通 Huitong (x 18)	t2.20
立山 *Lishan* (x 34)	道開 *Daokai* (x 19)	普眼菩薩 *Puyan pusa* (x 18)	t2.21
陶鏞 *Tao Yong* (x 33)	佛陀波利 *Fotuoboli* (x 18)	觀世音菩薩 *Guanshiyin pusa* (x 17)	t2.22
清了 *Qingliao* (x 32)	攝摩騰 *Shemoteng* (x 18)	性一 *Xingyi* (x 17)	t2.23
真可 *Zhenke* (x 31)	阿育王 *Ayu wang* (x 17)		t2.24

[a]The entry "unknown" is used for all names that the encoders were not able to identify. t2.25

similar are Buddhist temple gazetteers in this, for instance, to the corpus of letters and notebooks from the Song and Yuan dynasties[P]? Without a tagged corpus of notebooks this is difficult to answer, but it would be interesting to know if there is a genre specific "name density", which would allow for automated genre detection, or help to flag "eccentric compilations", which contain texts with unusual high or low amounts of names and dates.

Could the Dharma Drum gazetteer corpus be used as a training corpus for NER on Song-Yuan notebooks and letters? In principle yes, but with some caveats. Gazetteers generally are compilations of text from very different genres. This makes them interesting historical sources, but for a training corpus for 筆記 *biji* literature one must consider some adjustments. All verse passages (<l>s within <lg>s), for instance, should be removed from the gazetteer corpus, if it were to serve as training data for a corpus of *biji* literature (which consists almost exclusively of prose). Another feature that has to be removed from or disregarded in the gazetteer corpus is punctuation, as machine learning algorithms would latch on to that for feature recognition (if not specifically instructed not to) as has been shown for modern Chinese (Rao and Xun 2012: 18–19). Nevertheless the gazetteer corpus could certainly serve as next-of-kin for a corpus of Song-Yuan notebooks, or the corpus of official dynastic histories (正史 *zhengshi*) and allow benchmarking and aid digital analysis.