文獻標注與文史研究

祝平次

清華大學中國文學系副教授

「古籍全文資料庫的回顧與展望」工作坊

中央研究院數位文化中心
清華大學中國文學、科技部數位人文籌備小組

2014-06-05

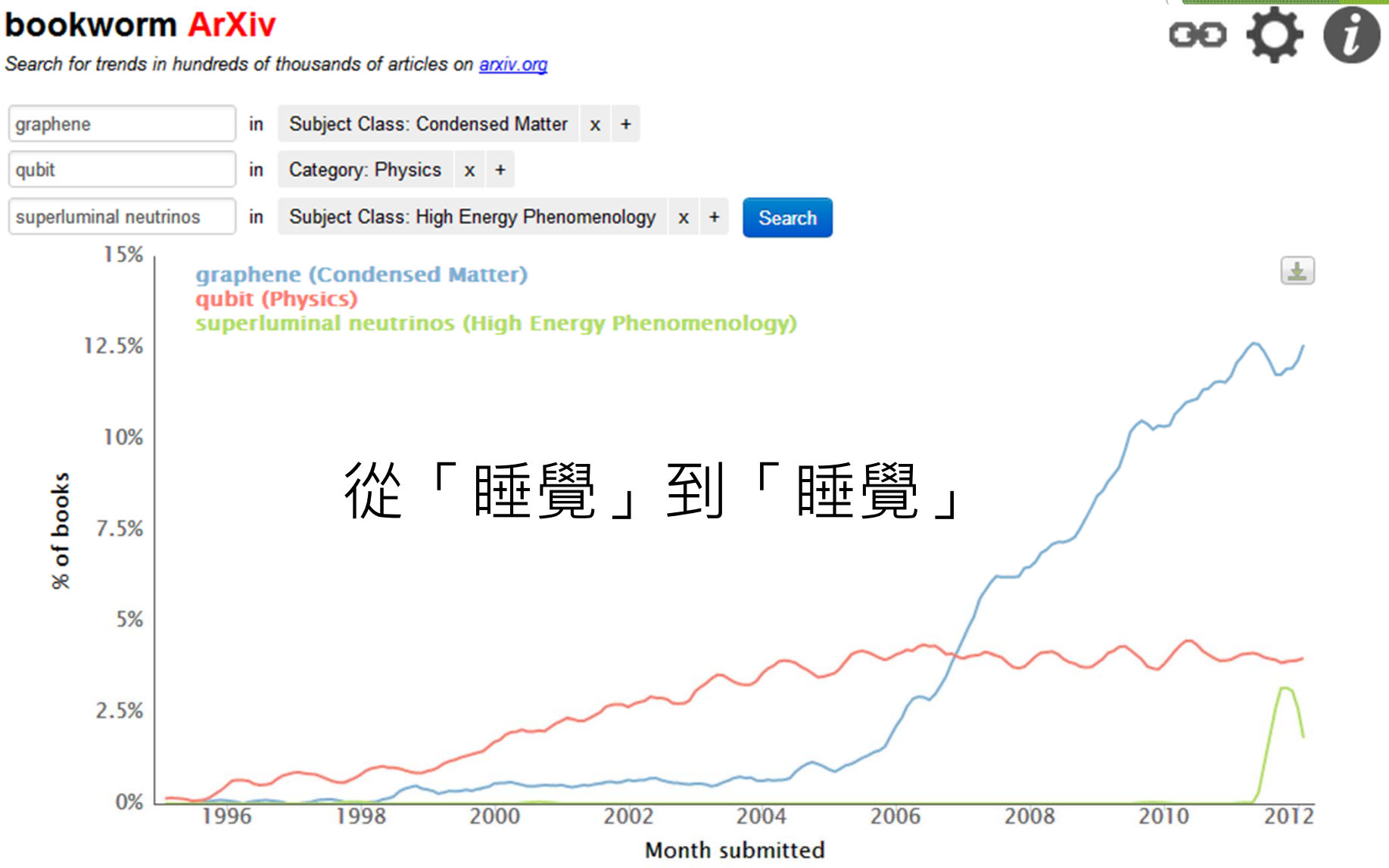
(部分投影片為萊登大學何浩洋所提供，特此誌謝)

報告大綱

- 目的：示範文獻標注以後對於文史研究的幫助(如何擴充全文資料的用途)
- 文獻標注的用處
- 標注的方法
- 配合CBDB的自動標注系統Markus
- 個案示範：黃榦《勉齋集》和陳淳《北溪大全集》：計量、GIS與社會網絡比較
- 結論

文獻標注的用處

利用大量文本分析特定詞彙不同時代的情況

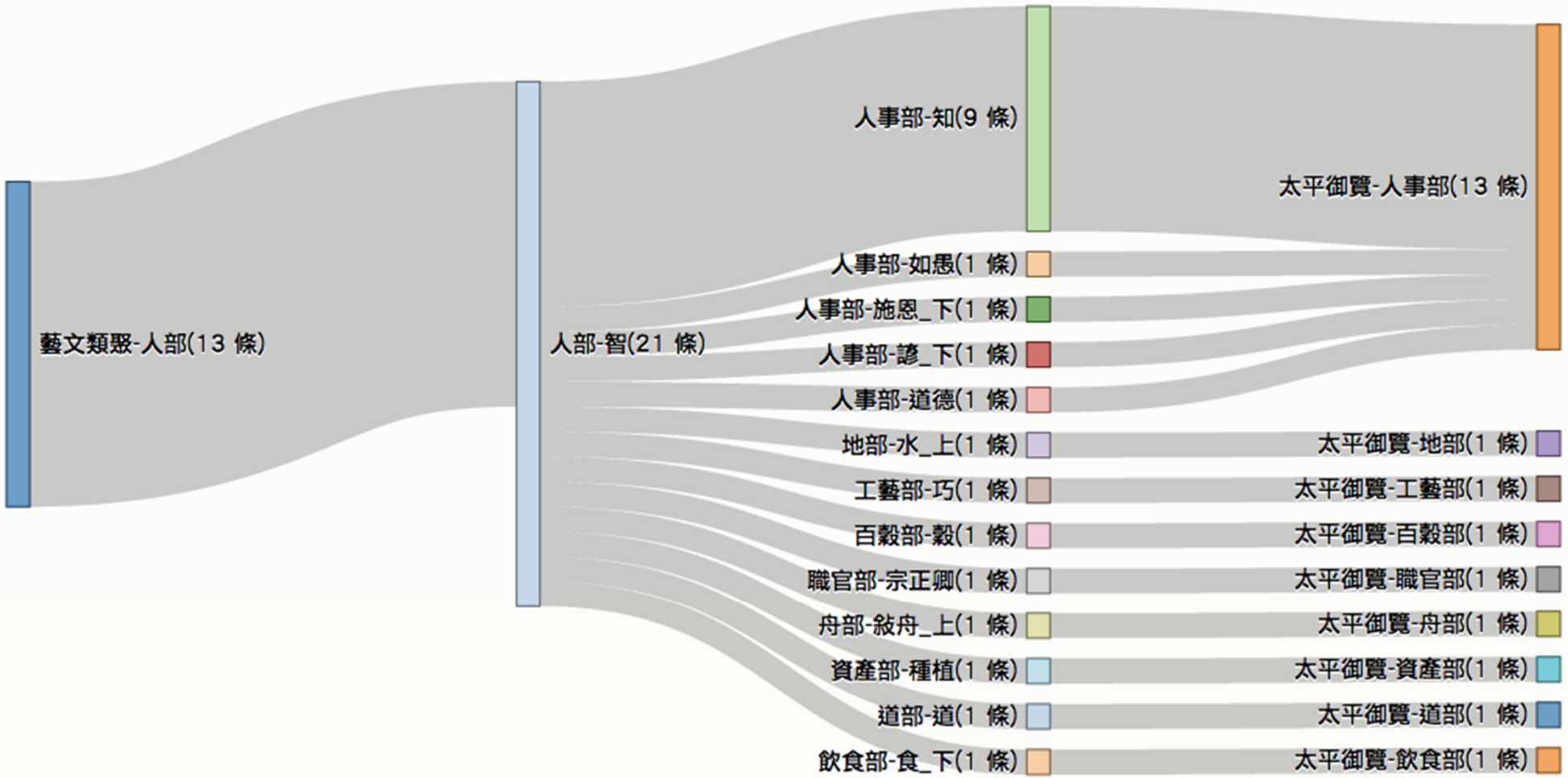


從「睡覺」到「睡覺」

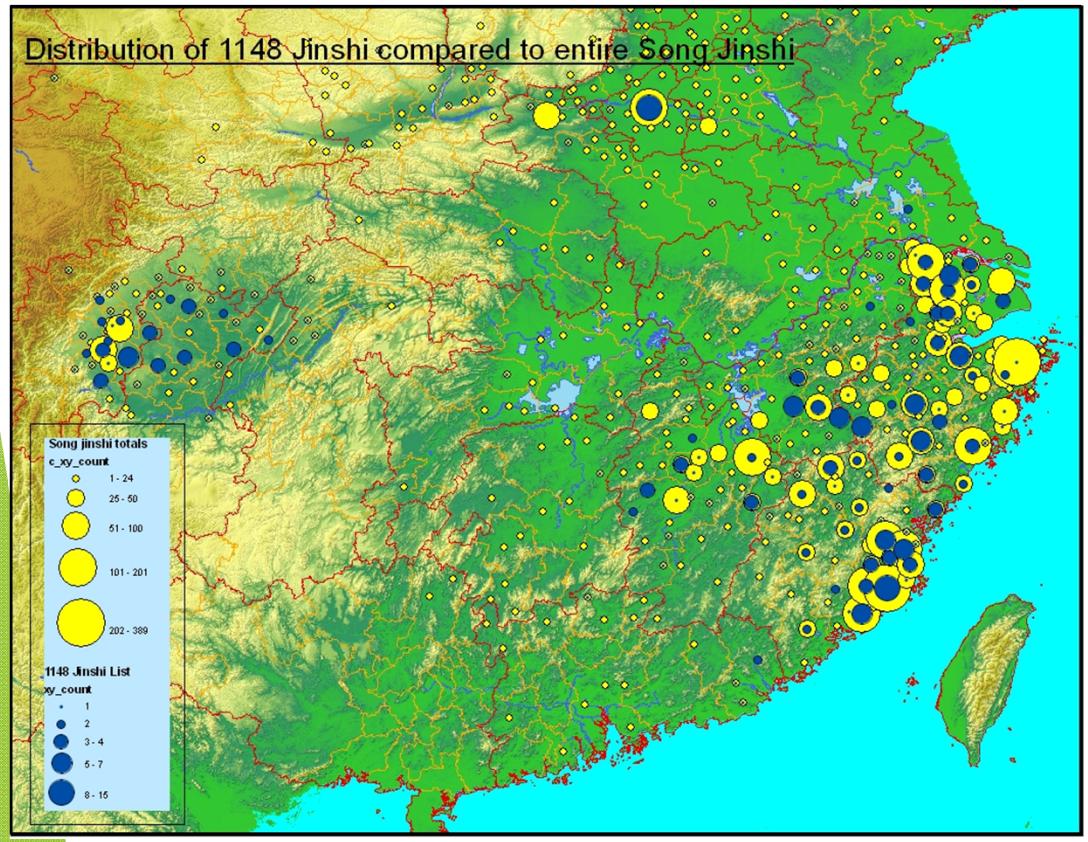
Culturomics (google n-gram viewer)

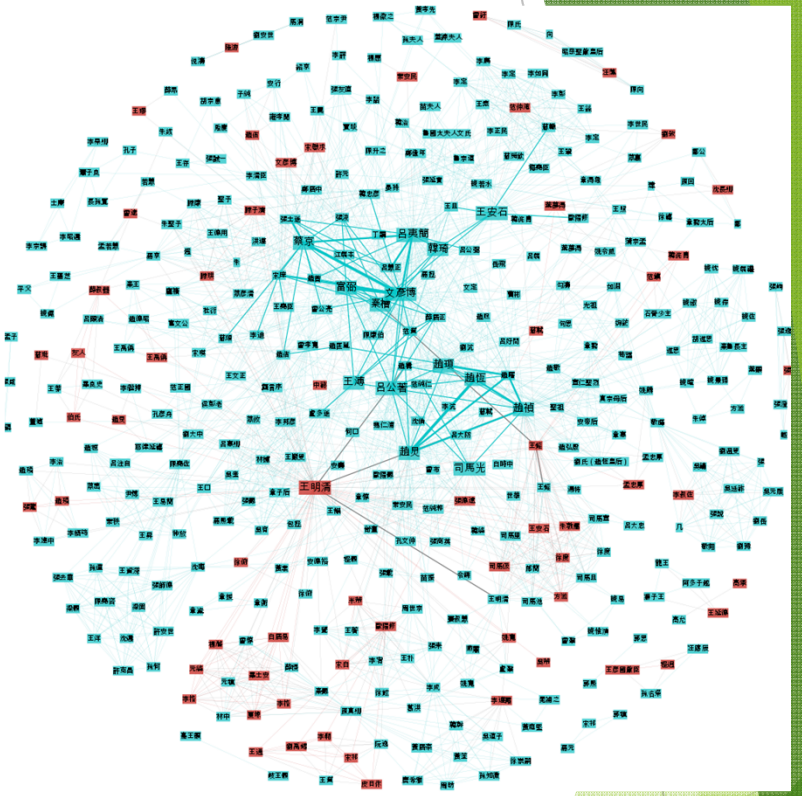
分析不同類書之間的相關性

◦ 統計



利用地理、人際網絡觀察中國歷代人物

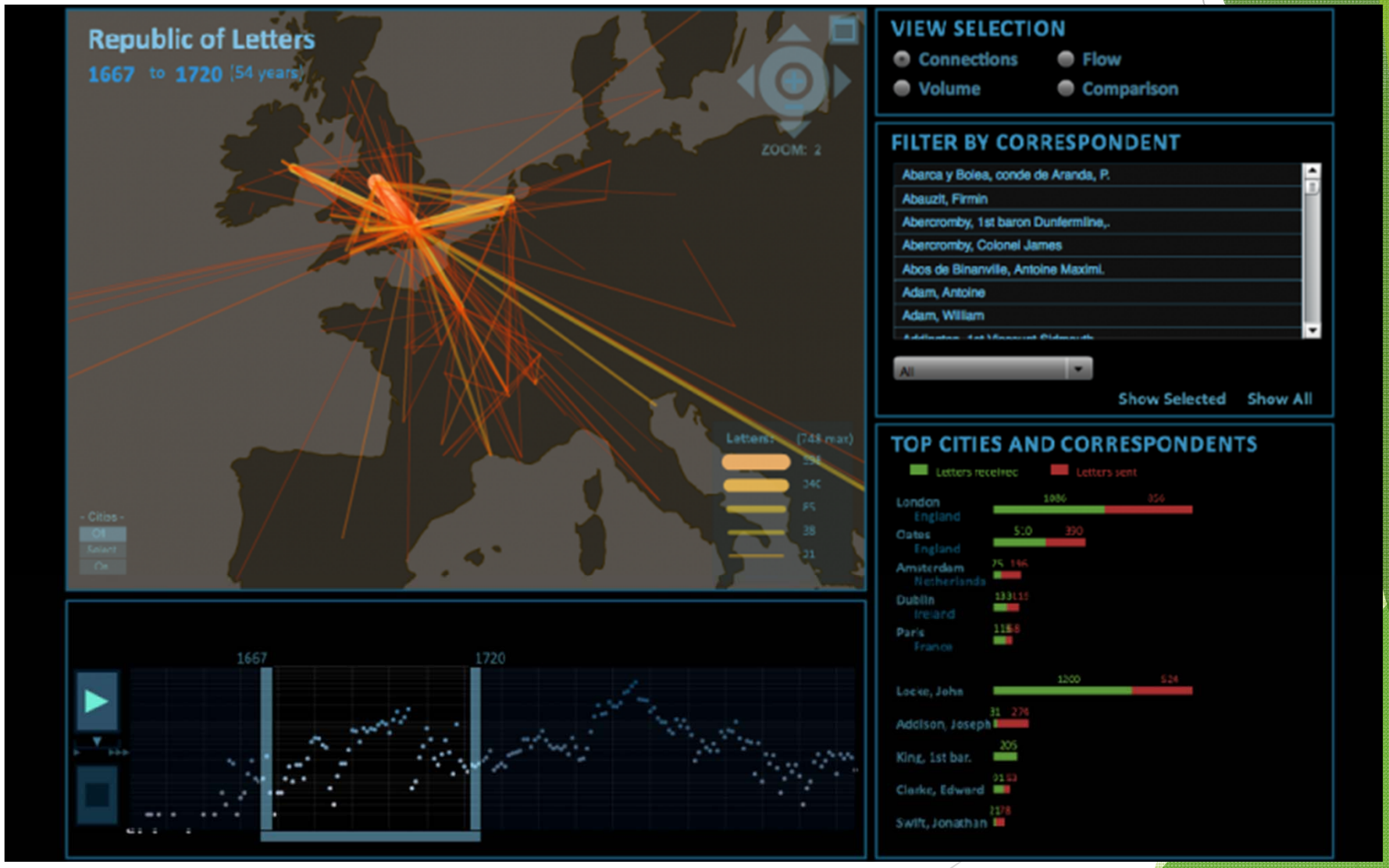




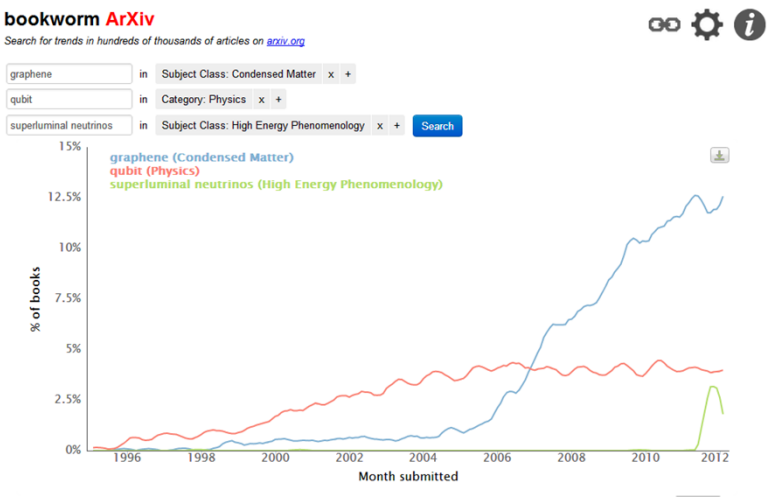
Communication and empire

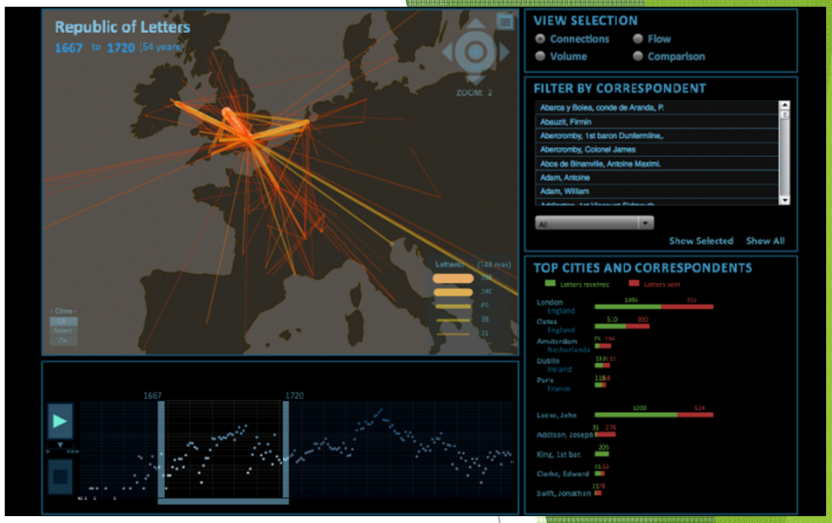
China biographical database (CBDB)

利用地理、時間、空間分析信件的流動

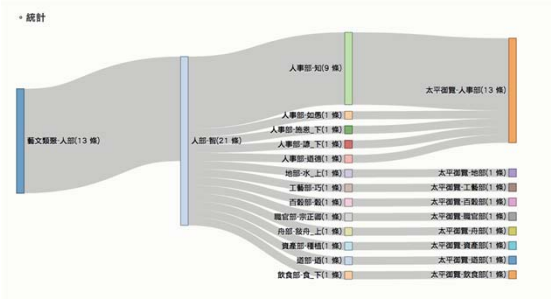


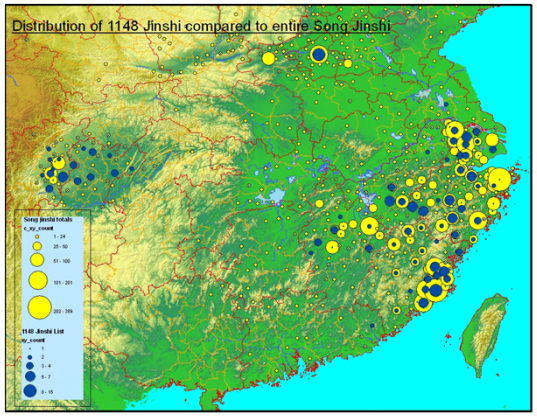
Republic of letters

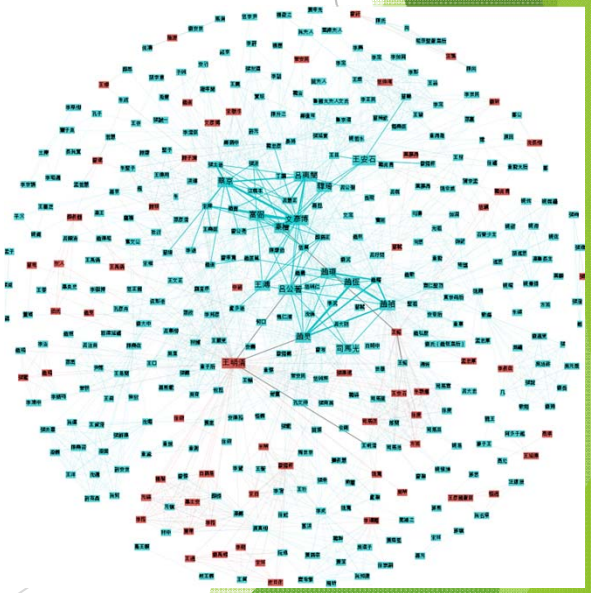

bookworm ArXiv
Search for trends in hundreds of thousands of articles on arxiv.org
graphene in Subject Class: Condensed Matter
qubit in Category: Physics
superluminal neutrinos in Subject Class: High Energy Phenomenology
Search
graphene (Condensed Matter)
qubit (Physics)
superluminal neutrinos (High Energy Phenomenology)
% of books
Month submitted


Republic of Letters
1667 to 1720 (54 years)
VIEW SELECTION
Connections
Flow
Volume
Comparison
FILTER BY CORRESPONDENT
Show Selected
Show All
TOP CITIES AND CORRESPONDENTS

數據


統計


Distribution of 1148 Jinshi compared to entire Song Jinshi



宋會要輯稿

文獻集: 宋會要輯稿

文件檢索: 邵雍

測試版

最近幾次檢索

邵雍	3
婺州	208
{TM:****} .all	18878
.all	80396

呈現模式: 檢索結果 | 逐篇檢視 | 詞頻與全文 | 列印全文

排列方式: 檔名順序 | 年代

找到筆數: 3　　頁次: 1

相關人名：邵雍生平・蘇洵生平・魏漢津生平・韓琦生平・劉光祖生平・常安民生平

相關地名：霸州・燕・潁州・衛州・河南・共城・文安

依 df降冪排列 | 依 t→q降冪排列

人名

Term t		(tf)	df	t→q
邵雍	生平	(3)	2	1.000
徐積	生平	(2)	1	0.250
蘇洵	生平	(3)	1	0.167
李堅	生平	(1)	1	0.091
劉光祖	生平	(1)	1	0.083
魏漢津	生平	(1)	1	0.083
常安民	生平	(1)	1	0.077
方時	生平	(1)	1	0.071
范祖禹	生平	(1)	1	0.0…
韓琦	生平	(1)	1	

更多…

依 df降冪排列 | 依 t→q降冪排

地名

Term t	(tf)	df	t→
共城	(1)	1	0.083
文安	(2)	1	0.019
潁州	(1)	1	0.018
霸州	(1)	1	0.015
衛州	(1)	1	0.011
河南	(1)	1	0.004
燕	(1)	1	0.002

依 df降冪排列 | 依 t→q降冪排

官名

Term t	(tf)		
太常	(3)		
團練推官	(1)		
將作監主簿	(1)	1	0.0…
宣德郎	(1)	1	0.011
光祿寺丞	(1)	1	0.009
著作郎	(1)	1	0.007
縣主簿	(2)	1	0.006
推官	(1)	1	0.005

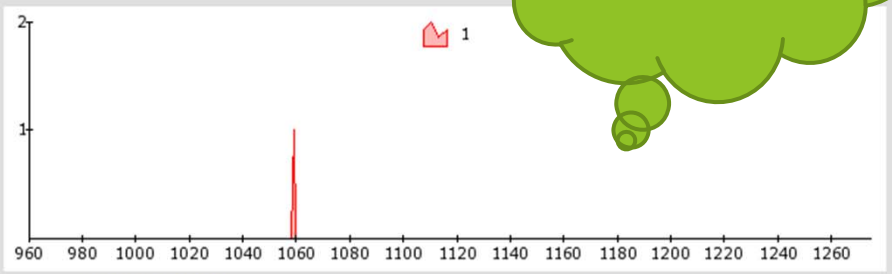
相關人名
詞頻分布

相關地名
詞頻分布

相關官名
詞頻分布

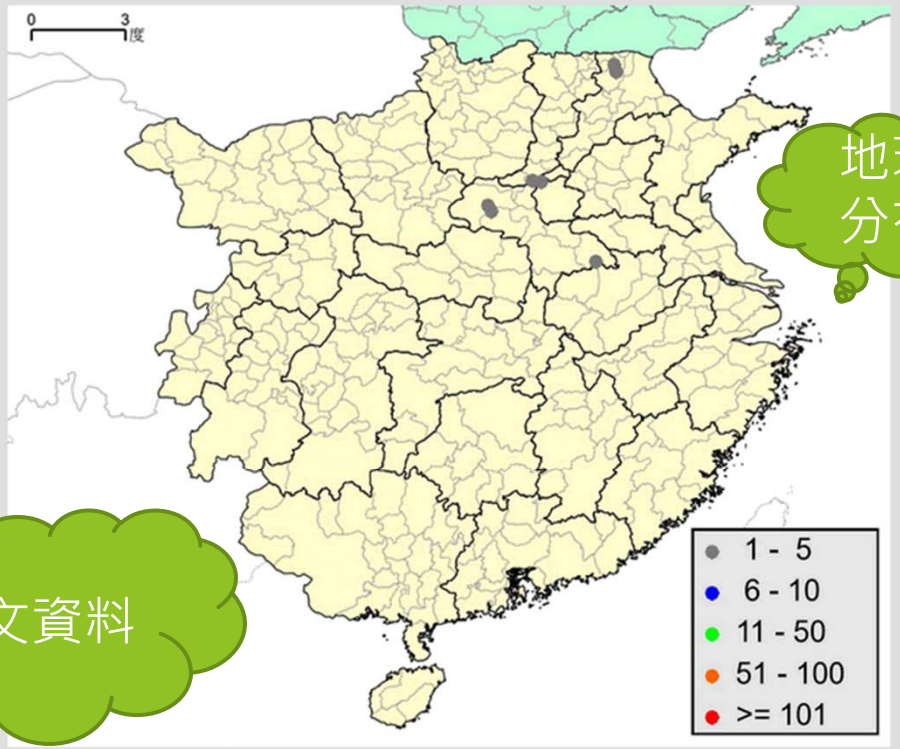
全文資料

年代分布



檢索結果分佈圖（此圖僅計入含有年代資訊的文件；橫軸:西元年，縱軸:文件數量）

地理
分布



地名文件數量分佈圖（此圖僅計入含有地名資訊的文件）

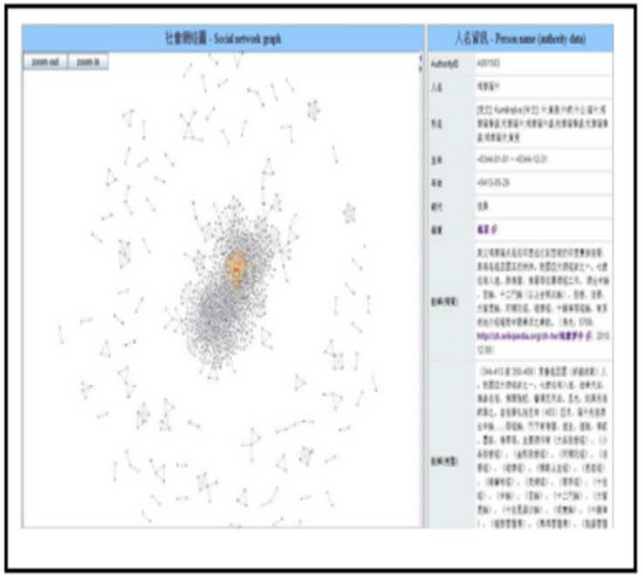
Buddhist Biographies

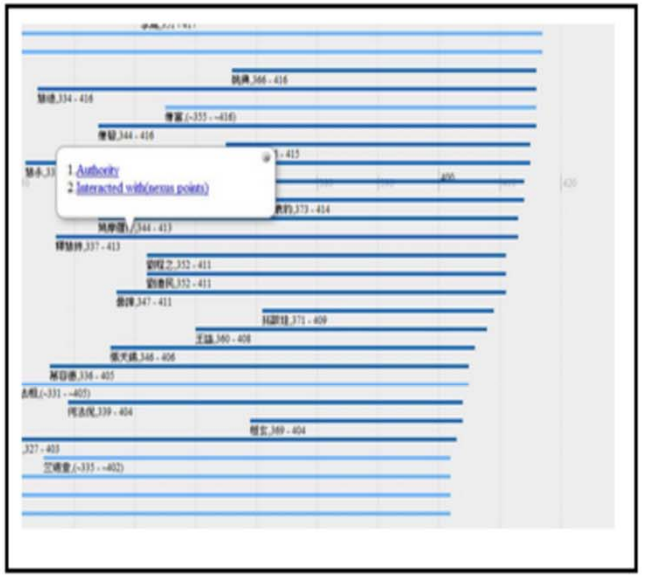
One Dataset

-

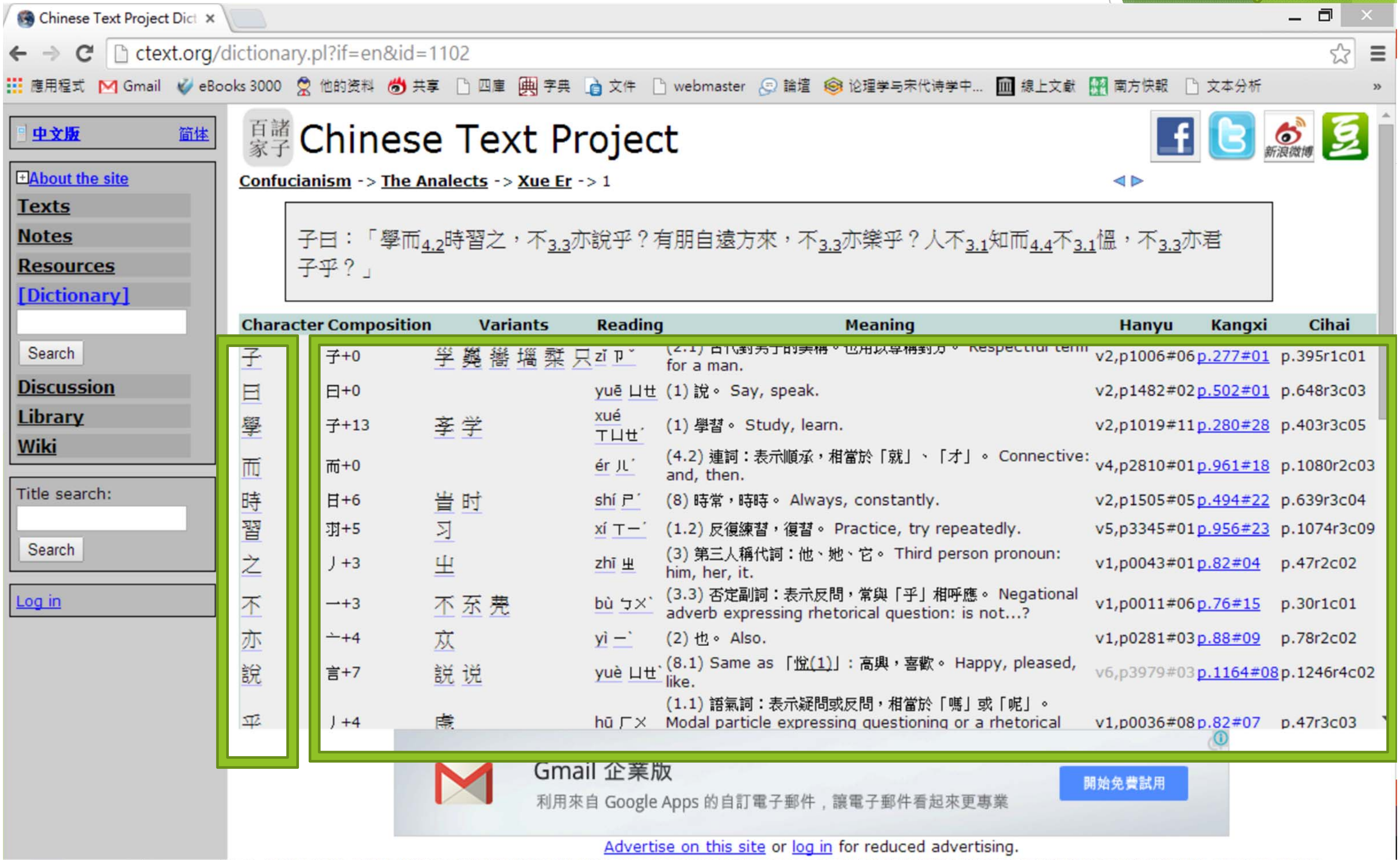
Three Views

中文





文獻標注的例子


Chinese Text Project Dict
ctext.org/dictionary.pl?if=en&id=1102
應用程式 Gmail eBooks 3000 他的资料 共享 四庫 字典 文件 webmaster 論壇 论理学与宋代诗学中... 線上文獻 南方快報 文本分析
中文版 简体
About the site
Texts
Notes
Resources
[Dictionary]
Search
Discussion
Library
Wiki
Title search:
Search
Log in
百諸家子 Chinese Text Project
Confucianism -> The Analects -> Xue Er -> 1
子曰：「學而4.2時習之，不3.3亦說乎？有朋自遠方來，不3.3亦樂乎？人不3.1知而4.4不3.1慍，不3.3亦君子乎？」
Character Composition Variants Reading Meaning Hanyu Kangxi Cihai
子 子+0 zǐ ㄗˇ for a man. v2,p1006#06 p.277#01 p.395r1c01
曰 日+0 yuē ㄩㄝ (1) 說。 Say, speak. v2,p1482#02 p.502#01 p.648r3c03
學 子+13 xué ㄒㄩㄝˊ (1) 學習。 Study, learn. v2,p1019#11 p.280#28 p.403r3c05
而 而+0 ér ㄦˊ (4.2) 連詞：表示順承，相當於「就」、「才」。 Connective: and, then. v4,p2810#01 p.961#18 p.1080r2c03
時 日+6 shí ㄕˊ (8) 時常，時時。 Always, constantly. v2,p1505#05 p.494#22 p.639r3c04
習 羽+5 xí ㄒㄧˊ (1.2) 反復練習，復習。 Practice, try repeatedly. v5,p3345#01 p.956#23 p.1074r3c09
之 丿+3 zhī ㄓ (3) 第三人稱代詞：他、她、它。 Third person pronoun: him, her, it. v1,p0043#01 p.82#04 p.47r2c02
不 一+3 bù ㄅㄨˋ (3.3) 否定副詞：表示反問，常與「乎」相呼應。 Negational adverb expressing rhetorical question: is not...? v1,p0011#06 p.76#15 p.30r1c01
亦 亠+4 yì ㄧˋ (2) 也。 Also. v1,p0281#03 p.88#09 p.78r2c02
說 言+7 yuè ㄩㄝˋ (8.1) Same as 「悅(1)」：高興，喜歡。 Happy, pleased, like. v6,p3979#03 p.1164#08 p.1246r4c02
乎 丿+4 hū ㄏㄨ (1.1) 語氣詞：表示疑問或反問，相當於「嗎」或「呢」。 Modal particle expressing questioning or a rhetorical v1,p0036#08 p.82#07 p.47r3c03
Gmail 企業版
利用來自 Google Apps 的自訂電子郵件，讓電子郵件看起來更專業
開始免費試用
Advertise on this site or log in for reduced advertising.

文獻標注的用處

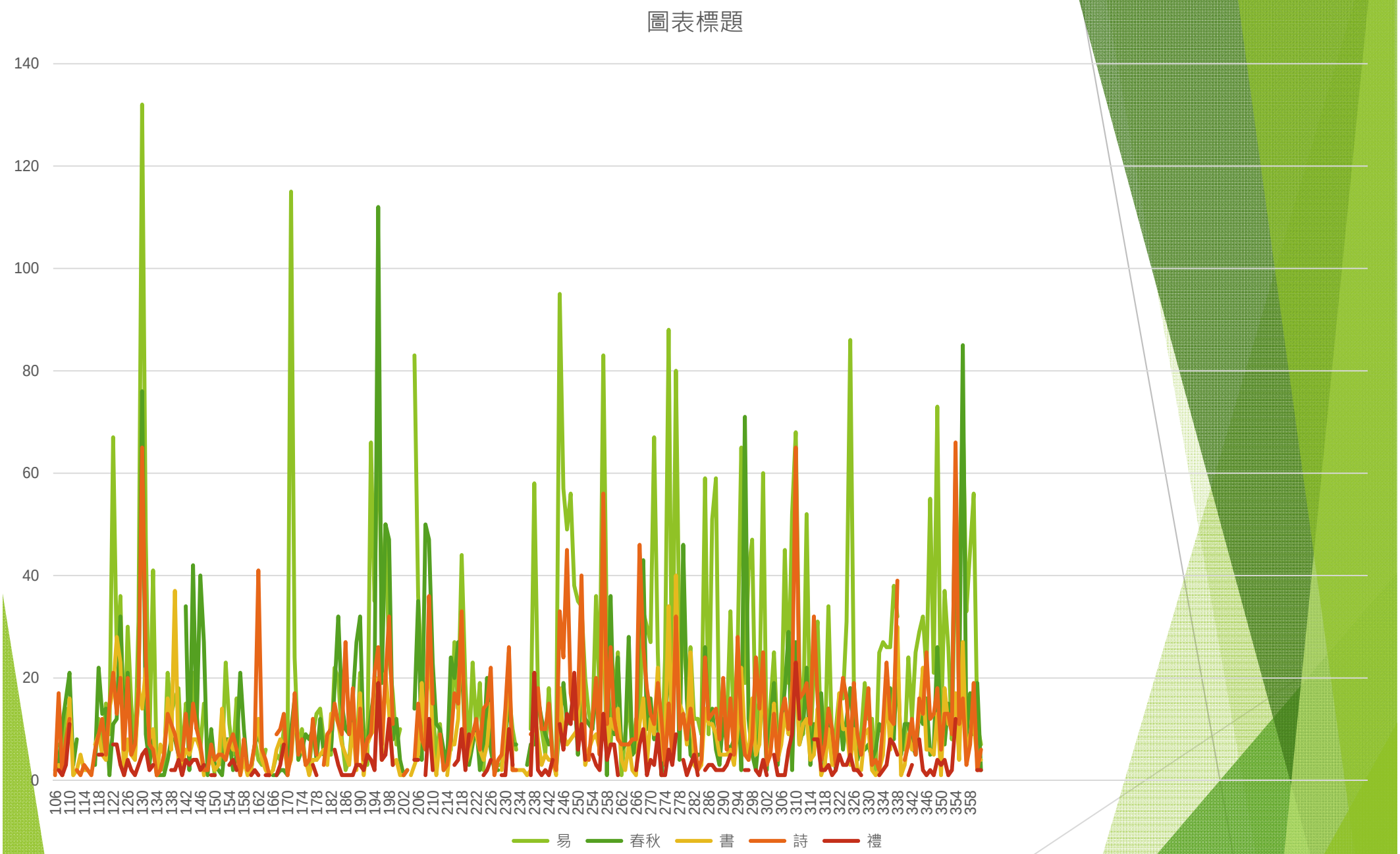
- 可以用不同的方式審視(有研究的意味)與展示(分享知識的意味)文獻
  - 包括轉化成其它媒體；
  - 包括顯題化文獻的某些特性；
  - 成為「數」據。
- 與其它數據的串連
  - 其為文獻也不同，其為數據也則一；
  - 同類文獻的整合(四庫全書、類書)；
  - 跨文類的整合(Google圖片搜尋、將音樂轉化成圖示)。
- 與電腦進行互動(人機互動)
  - 數位工具的產生：時間線圖、族譜
  - 對於人工研究過程的省視：年譜

可能的研究課題舉例

- 全宋文提到的文獻研究
  - 有多少還存在？
  - 有沒有時段前後的差別？
  - 不同文類的引用情形一樣不一樣？意味著什麼？
  - 不同作者(例如理學家、非理學)引用的情形一樣不一樣？意味著什麼？
  - 文獻彼此之間的關係(使用在同一個單位出現來建構文獻之間的關係)呈現什麼樣的狀況？意味著什麼？

書名	冊	ID1	ID1	篇ID	作者	卷數
祭法	1			qswi_001006	宋太祖	qswj_0010001
國風	1			qswi_001015	宋太祖	qswj_0010001
國風	1			qswi_001018	宋太祖	qswj_0010001
歸妹	1			qswi_001018	宋太祖	qswj_0010001
易	1			qswi_001018	宋太祖	qswj_0010001
三禮圖	1			qswi_001030	宋太祖	qswj_0010001
祠令	1			qswi_001066	宋太祖	qswj_0010002
重詳定刑統節文	1			qswi_001099	宋太祖	qswj_0010003
詩	1			qswi_001105	宋太祖	qswj_0010003
國風	1			qswi_001105	宋太祖	qswj_0010003
內則	1			qswi_001105	宋太祖	qswj_0010003
薤露	1			qswi_001106	宋太祖	qswj_0010003
禮	1			qswi_001106	宋太祖	qswj_0010003
書	1			qswi_001106	宋太祖	qswj_0010003

易6775	周官808	孝經391
春秋4538	坤777	史記360
詩4261	周禮729	尚書345
書3158	孟子655	坎341
禮1289	洪範540	震335
中庸1244	雅497	艮323
論語1105	記492	左氏322
乾1012	象401	孟321
大學1003	離397	禮記318
傳975	語394	頌295


圖表標題
140
120
100
80
60
40
20
0
106 110 114 118 122 126 130 134 138 142 146 150 154 158 162 166 170 174 178 182 186 190 194 198 202 206 210 214 218 222 226 230 234 238 242 246 250 254 258 262 266 270 274 278 282 286 290 294 298 302 306 310 314 318 322 326 330 334 338 342 346 350 354 358
易
春秋
書
詩
禮

易	6775	詩	4261
朱熹	405	朱熹	201
李綱	152	魏了翁	99
魏了翁	142	歐陽修	88
晁説之	132	周必大	72
李石	100	陳傅良	71
楊簡	97	劉克莊	68
楊萬里	95	晁説之	64
曾丰	92	司馬光	63
胡銓	92	薛季宣	60
黃裳	87	蘇軾	55
薛季宣	79	陳舜俞	50
陽枋	77	楊萬里	47
陳淳	74	真德秀	46
葉適	69	王安石	44
真德秀	68	蘇轍	43
劉克莊	66	周紫芝	40
蘇軾	61	葉適	39

全文資料是資料庫嗎？

- 全文資料的資料庫性質是隱性的。
- 以前的人就有隱性資料庫的概念；也就是分類的概念—篇章節段、分類、類書。

標注的方法

如何讓不會中文的人分辨出人名、地名

紹興年間紹興在紹興喝紹興酒

如何讓不會中文的人分辨出人名、地名

紹興年間紹興在紹興喝紹興酒

—

標注(Markup)
讓電腦了解文本

<nianhao>紹興</nianhao>年間
<name type="person"><family_name>紹
</family_name><given_name>興
</given_name></name>
在<name type="place">紹興</name>喝
<object type="alcohol">紹興酒</object>

過溫寄鞏縣宰吳秘丞 原注：皇祐元年。

- <head><date when="1049"/>
- <c>過</c><placeName><c>溫</c></placeName><c>寄</c><persName><placeName><c>鞏</c><c>縣</c></placeName><addName type="office"><c>宰</c></addName><surname><c>吳</c></surname><addName type="office"><c>秘</c><c>丞</c></addName></persName><hi>原注：<date notBefore="1049-02-05">皇祐元年</date>。</hi></head>

過溫寄鞏縣宰吳秘丞 原注：皇祐元年。

- <head><date when="1049"/>
- <hi>原注：<date notBefore="1049-02-05">皇祐元年</date>。</hi></head>
- 兩個時間標記：
  一個標詩的著成年代，
  一個標詩的較準確的年代。

過溫寄鞏縣宰吳秘丞 原注：皇祐元年。

- <placeName><c>溫</c></placeName>

- 一個地名標記

過溫寄鞏縣宰吳秘丞 原注：皇祐元年。

- <persName><addName type =“職”><placeName><c>鞏</c><c>縣 </c></placeName><addName type=“office”><c>宰 </c></addName></addName><surname><c>吳 </c></surname><addName type=“官"><c>秘 </c><c>丞</c></addName></persName>

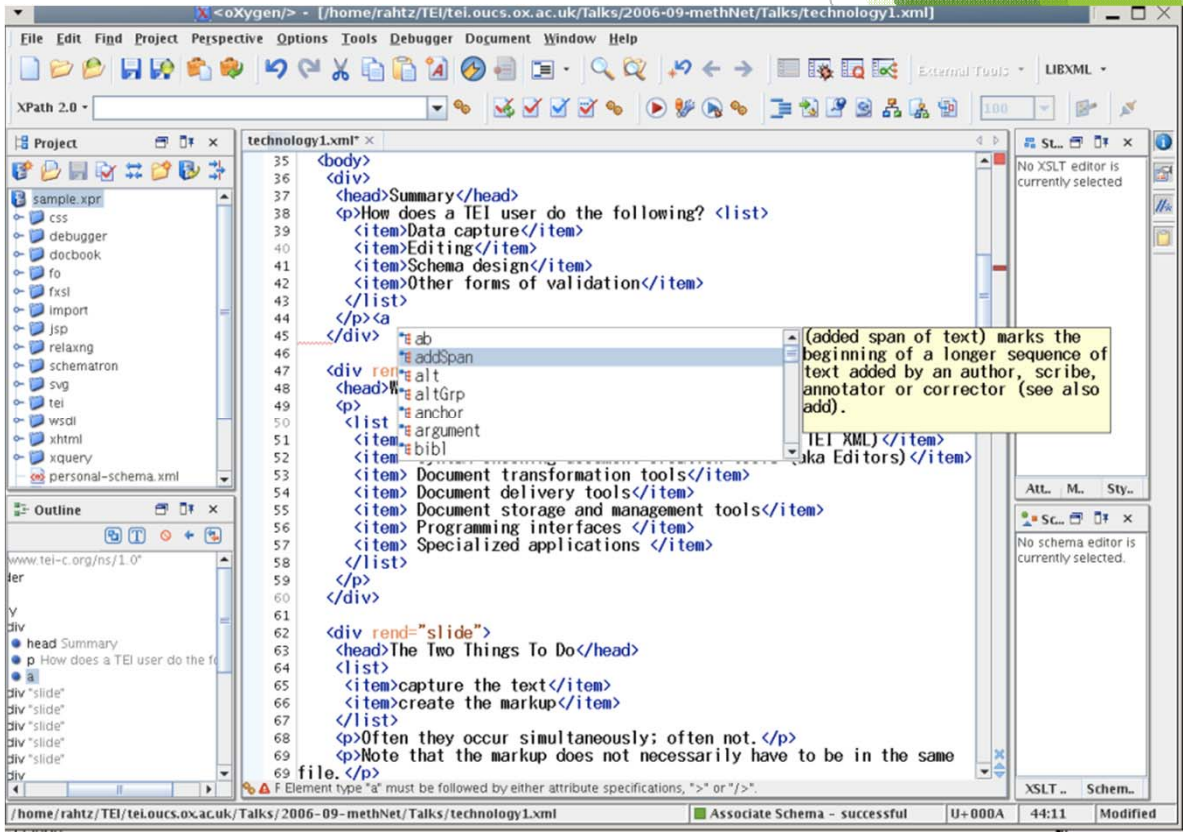
- 一個人名，其中還包括一個包括地名、職名的額外名、一個姓、一個官名的額外名。

XML的標註原則

- 標註不得交叉，必須要有個根標註。
- <root>
  - <Iloveyou>
    - <Ihateyou>
    - </Ihateyou>
  - </Iloveyou>
  - <home>
    - <room>Once upon a time</room>
    - <room>There was a prince</room>
  - </home>
- </root>

TEI - Text Encoding Initiative

- 第五版標準有1500 頁的說明
- 《佛教傳記文學視覺化平台》標記工作手冊(法鼓山)
- 標注工具：
  Oxygen XML editor

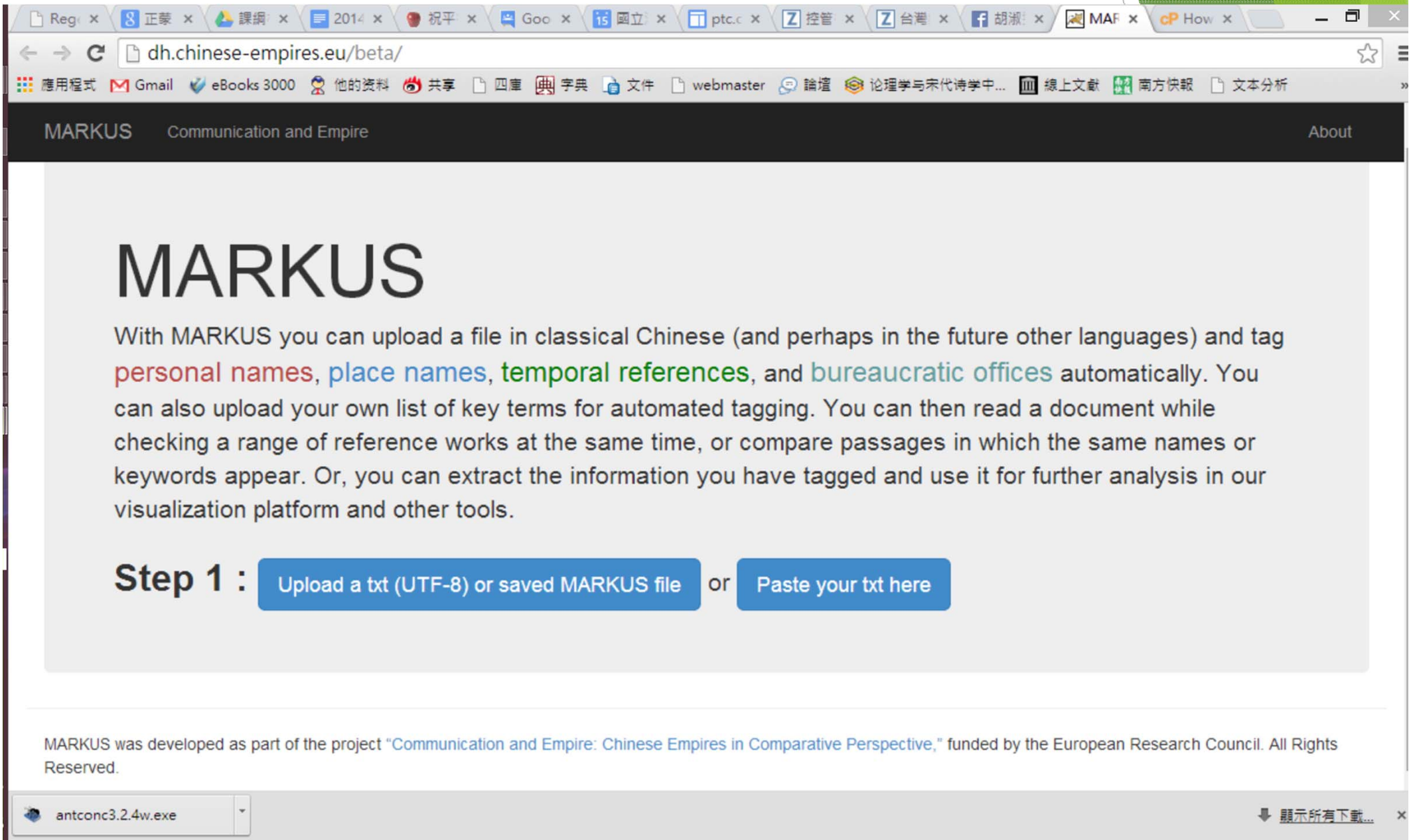


自動標注的必要性 (程式設計)

- 標注很有用，
  但人工標注很繁瑣，
  費時費力，
  很難進行。

配合CBDB的自動標注系統
Markus

Markus


dh.chinese-empires.eu/beta/
MARKUS
Communication and Empire
About
MARKUS
With MARKUS you can upload a file in classical Chinese (and perhaps in the future other languages) and tag personal names, place names, temporal references, and bureaucratic offices automatically. You can also upload your own list of key terms for automated tagging. You can then read a document while checking a range of reference works at the same time, or compare passages in which the same names or keywords appear. Or, you can extract the information you have tagged and use it for further analysis in our visualization platform and other tools.
Step 1 :
Upload a txt (UTF-8) or saved MARKUS file
or
Paste your txt here
MARKUS was developed as part of the project "Communication and Empire: Chinese Empires in Comparative Perspective," funded by the European Research Council. All Rights Reserved.
antconc3.2.4w.exe

Markus
(Mark us)

- 解決人工標注的問題：讓電腦進行自動標注。
- 利用CBDB(China Biography Database 哈佛大學、北京大學、史語所合作的中國歷代人物傳記資料庫)來進行
  人名、地名、官名、時間名
  的標註

Markus
http://dh.chinese-empires.eu/beta/

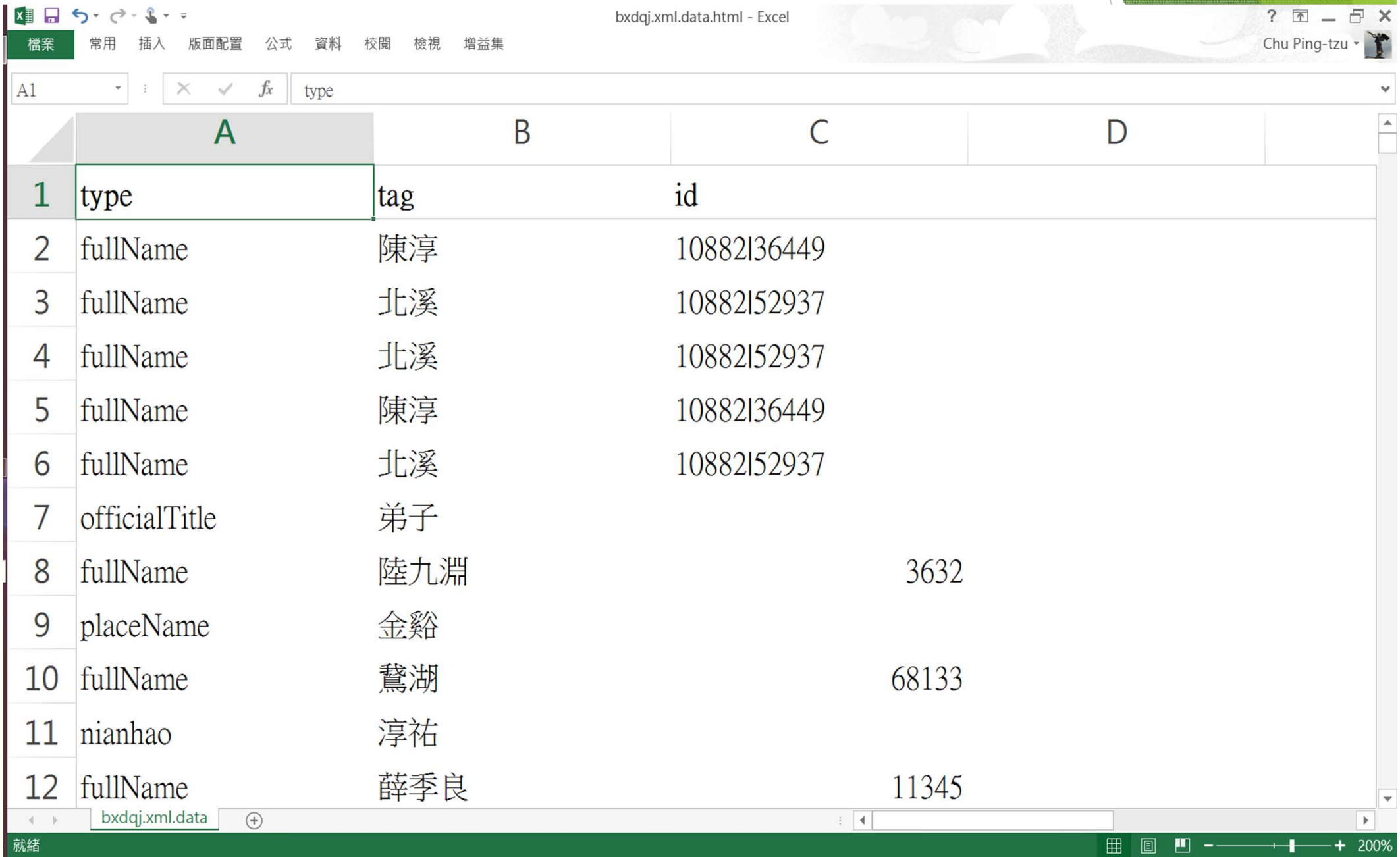
在標注之後

個案示範：
黃榦《勉齋集》
陳淳《北溪大全集》

儲存結果

陳淳			黃榦		
Type	Total	Unique	Unique	Total	Type
fullName	1018	329	624	1638	fullName
partialName	41	11	13	23	partialName
placeName	604	230	496	2072	placeName
officialTitle	651	162	436	2010	officialTitle
nianhao	460	75	98	527	nianhao

顯示結果



bxdqj.xml.data.html - Excel

檔案 常用 插入 版面配置 公式 資料 校閱 檢視 增益集

Chu Ping-tzu

A1 type

	A	B	C	D
1	type	tag	id	
2	fullName	陳淳	10882l36449	
3	fullName	北溪	10882l52937	
4	fullName	北溪	10882l52937	
5	fullName	陳淳	10882l36449	
6	fullName	北溪	10882l52937	
7	officialTitle	弟子		
8	fullName	陸九淵	3632	
9	placeName	金谿		
10	fullName	鵝湖	68133	
11	nianhao	淳祐		
12	fullName	薛季良	11345	

bxdqj.xml.data

就緒 200%

與CBDB連結

計量、GIS與社會網絡比較

比較顯示結果

列標籤	計數 - tag
10882	181
3257	73
43538	44
7164	16
10892	9
504	8
25123	7
19162	6
46581	5
42981	5
51981	5
49517	4
42758	4
28168	4
52892	3
11134	3
48607	3
1216	3
52959	3
20041	3
53695	3
22959	3
37813	3
42755	2
40992	2
47602	2
11871	2
42702	2
12165	2
44128	2
14421	2
3105	2
15296	2

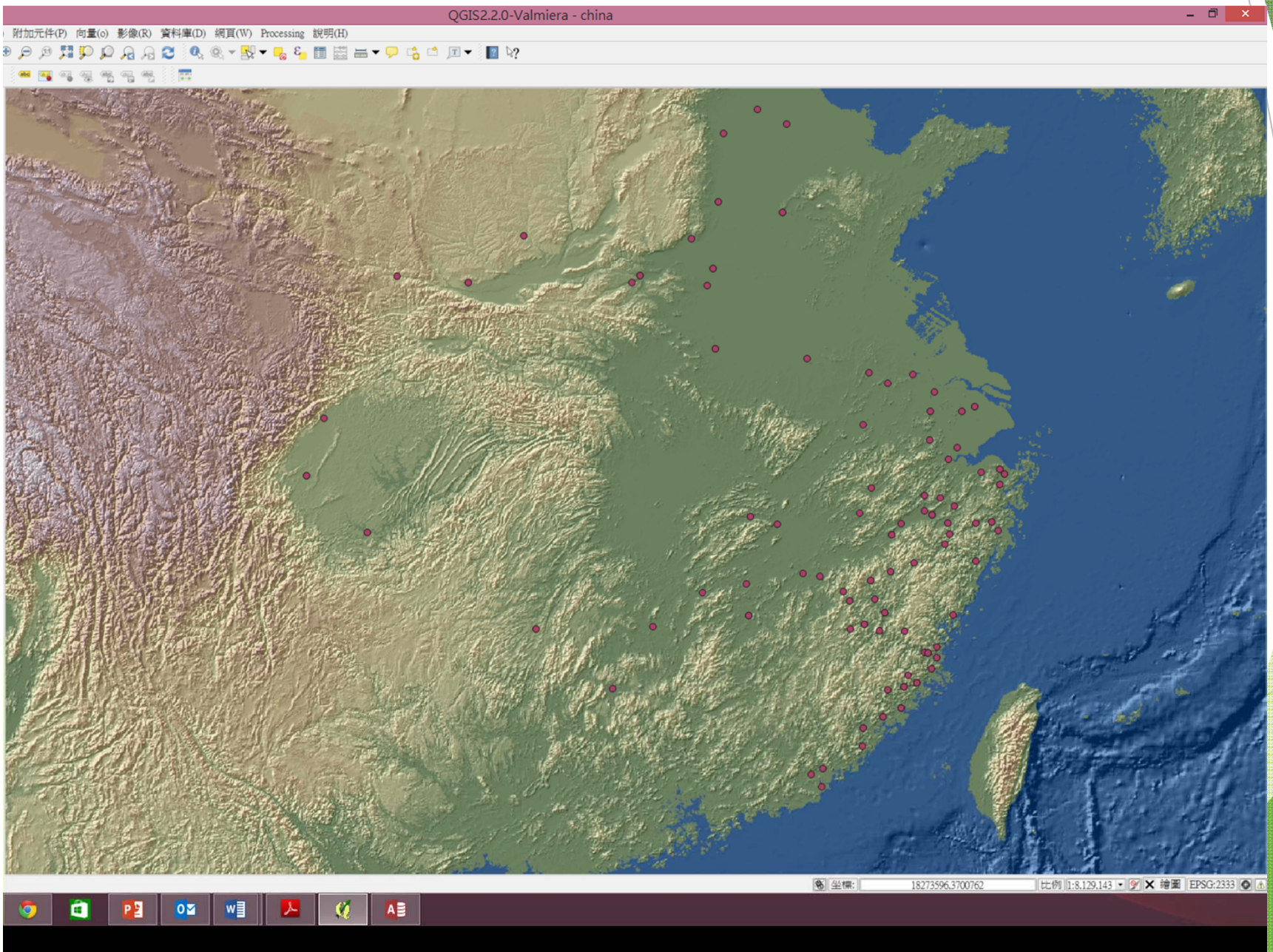
列標籤	計數 - tag
10882	181
北溪	124
安卿	3
陳淳	54
3257	73
文公	20
朱子	1
朱文公	6
考亭	6
晦翁	18
晦庵	7
紫陽	15
43538	44
伯澡	1
陳伯澡	40
陳沂	3
7164	16
南軒	15
南軒先生	1
10892	9
陳宓	4
復齋	5
504	8
濂溪	7
濂溪先生	1
25123	7
廖子晦	3
槎溪	4
19162	6
李公晦	6
46581	5
郭子從	5
42981	5
一之	2

列標籤	計數 - tag
11134	199
26991	36
7123	35
3257	26
20027	11
26590	11
11127	11
11827	10
42629	10
12408	9
38978	8
10892	7
14600	7
25535	7
40880	6
40723	6
7164	6
11904	6
51938	6
42031	5
30454	5
2129	5
22389	5
37056	5
22487	5
49814	5
22524	5
3597	5
44600	4
16209	4
51319	4
7055	4
44128	4
10304	4
47842	4
12053	4
16039	4

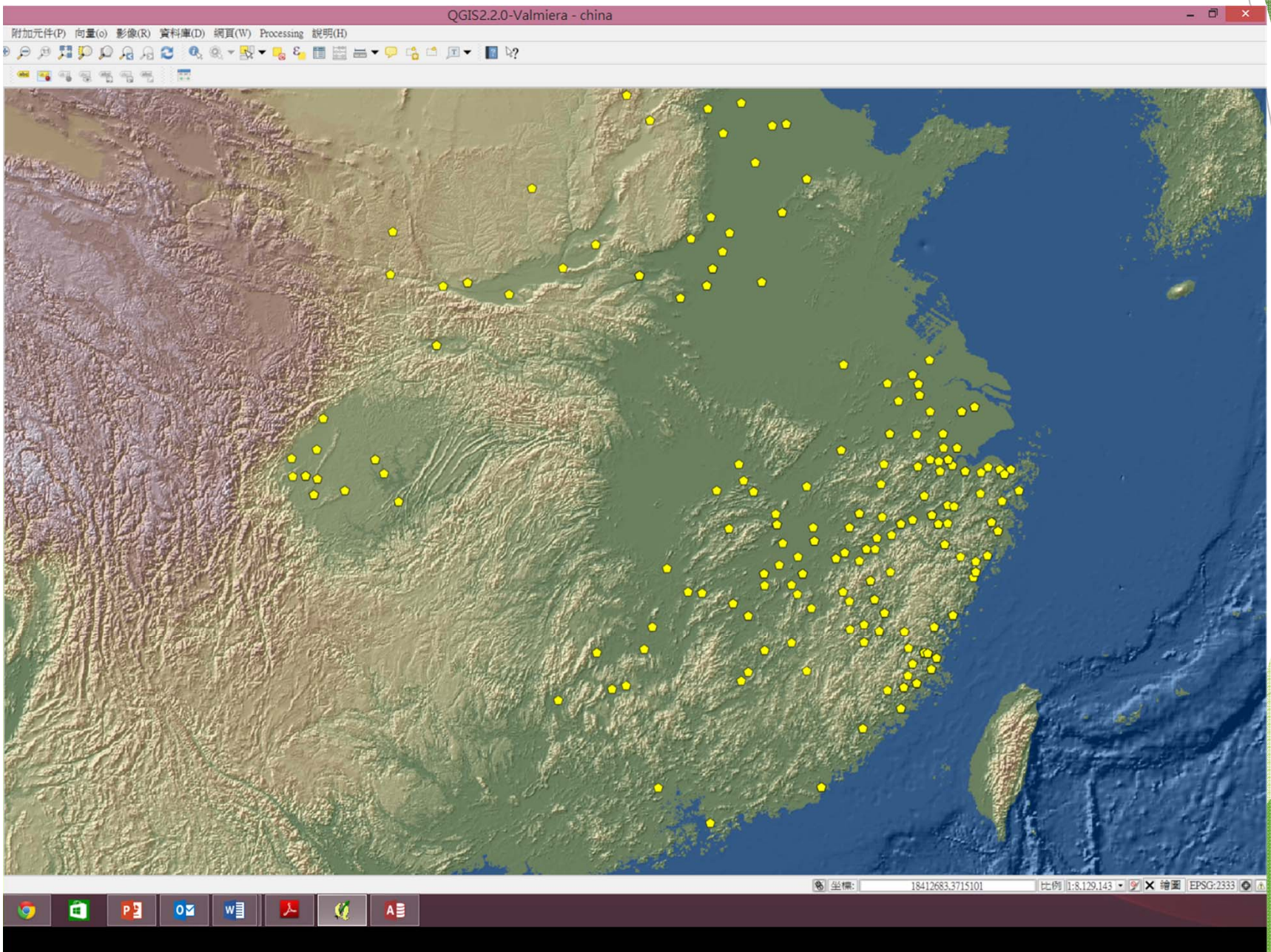
列標籤	計數 - tag
11134	199
直卿	2
勉齋	85
黃榦	112
26991	36
陳安國	36
7123	35
陳安節	35
3257	26
朱文公	10
朱熹	1
考亭	9
晦翁	4
紫陽	1
遯翁	1
20027	11
李貫之	10
李道傳	1
26590	11
伯量	1
胡伯量	10
11127	11
潘謙之	11
11827	10
屏山	10
42629	10
存齋	6
存齋先生	1
林公度	3
12408	9
張潛	9
38978	8
楊志仁	8
10892	7
陳宓	1
陳師復	6
14600	7

陳淳 | 陳淳人名表 | 陳淳人名 | 工作表14 | 陳淳官名表 | 陳淳官名 | 陳淳年號 | 陳淳地名 | 工作表3

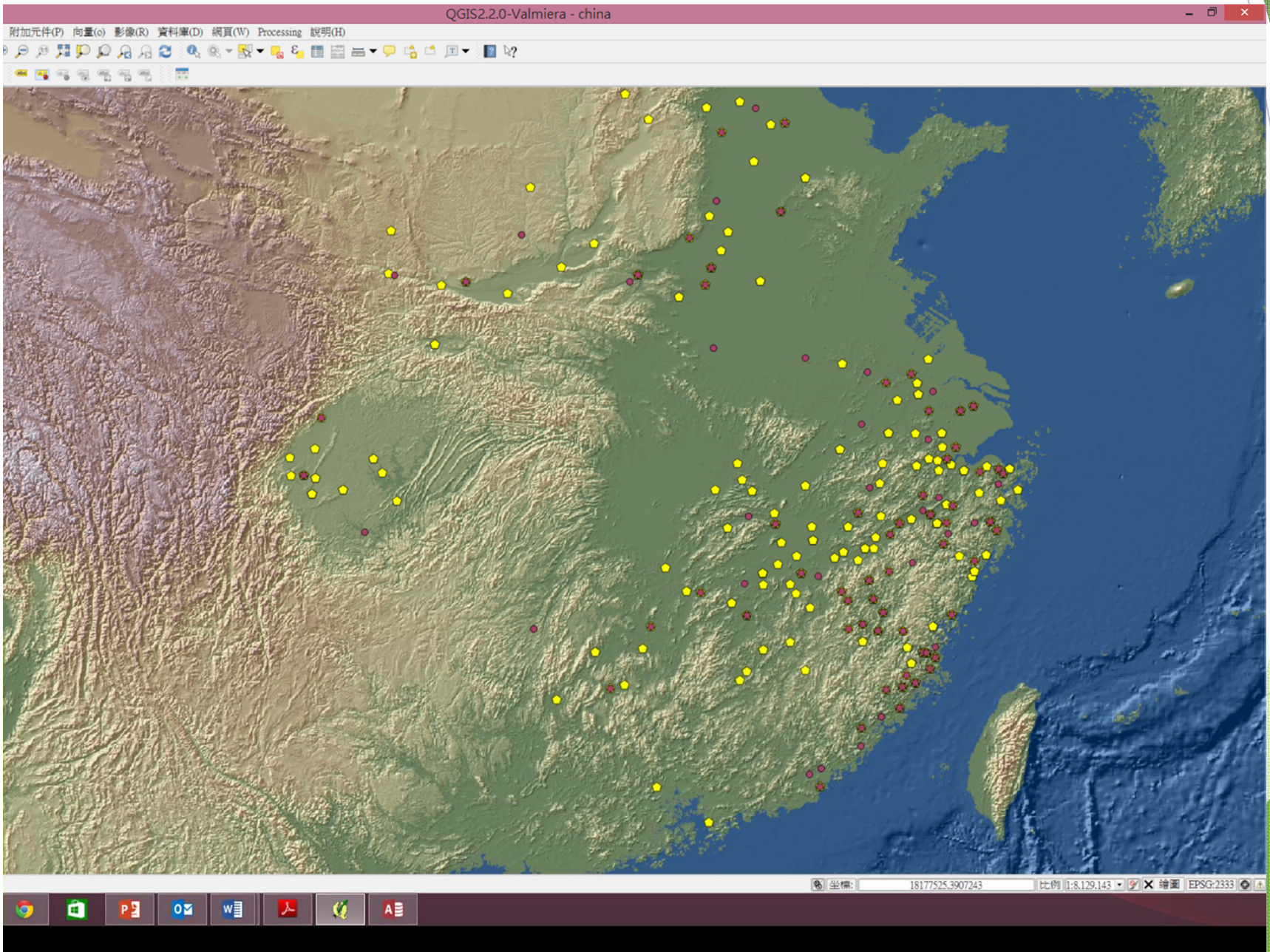
陳淳北溪大全集中人名的地理分布



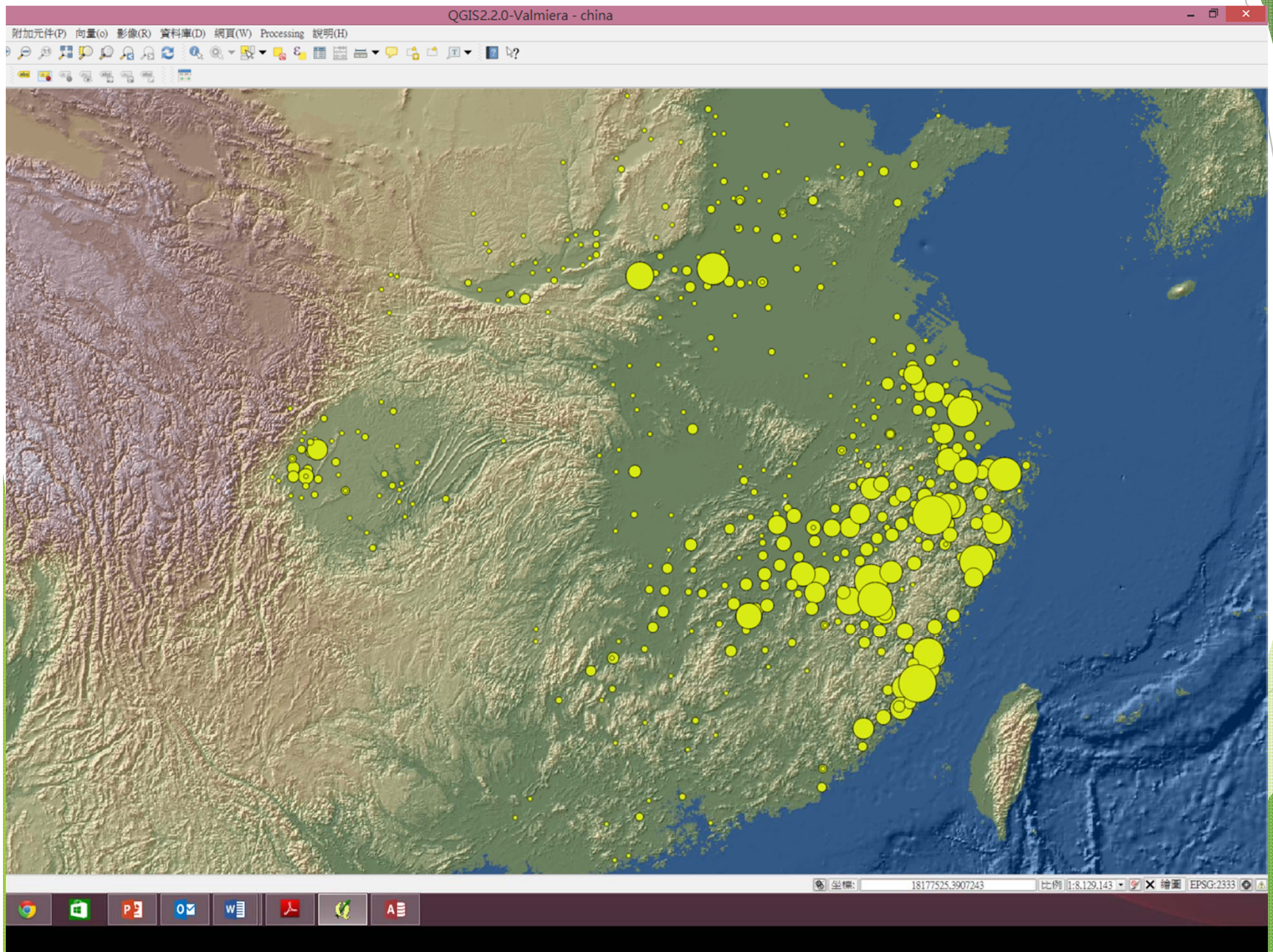
黃榦勉齋集中人名的地理分布



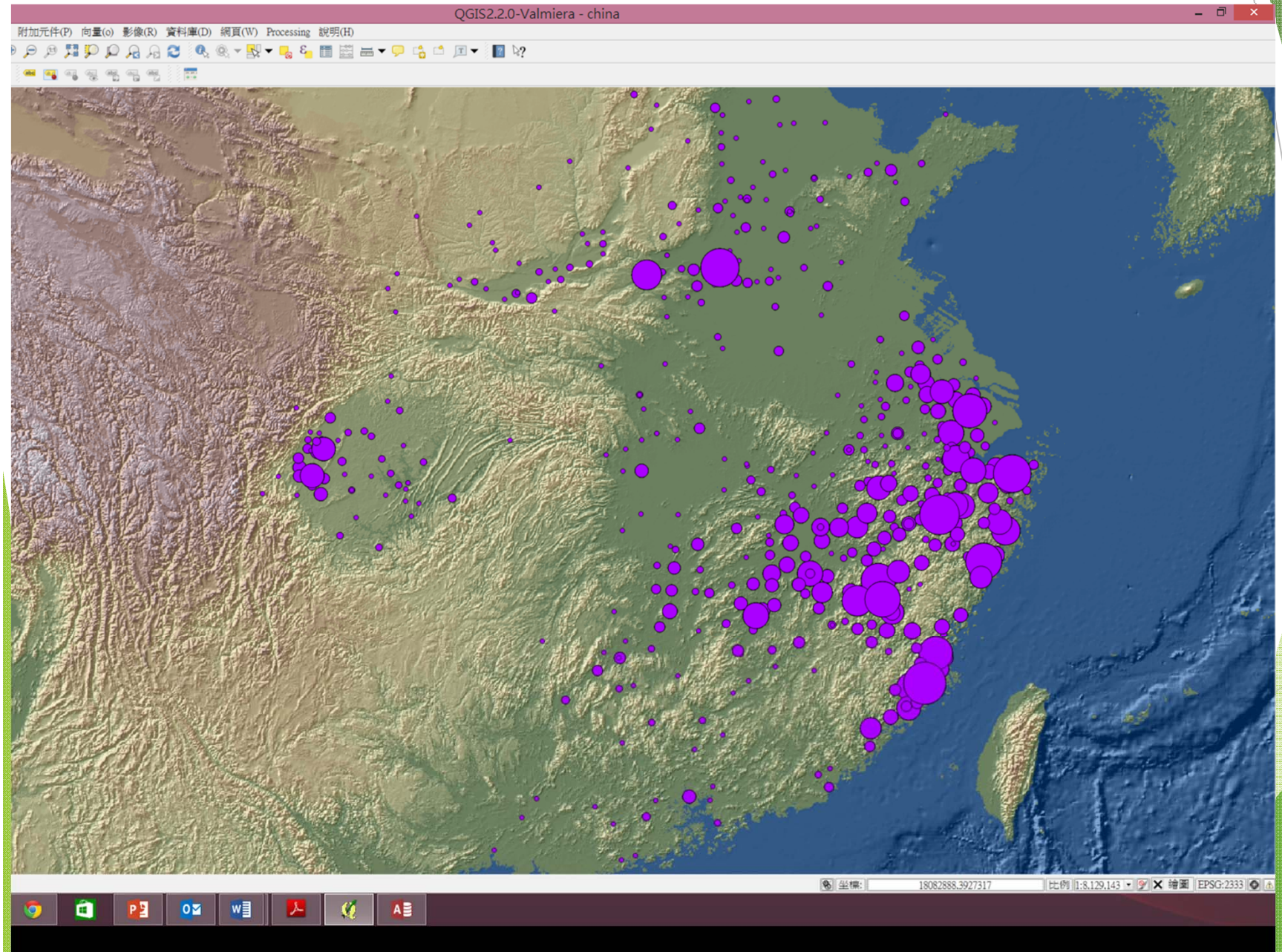
集中人名地理分布比較


QGIS2.2.0-Valmiera - china
附加元件(P) 向量(o) 影像(R) 資料庫(D) 網頁(W) Processing 說明(H)
18177525,3907243
比例 1:8,129,143
EPSG:2333

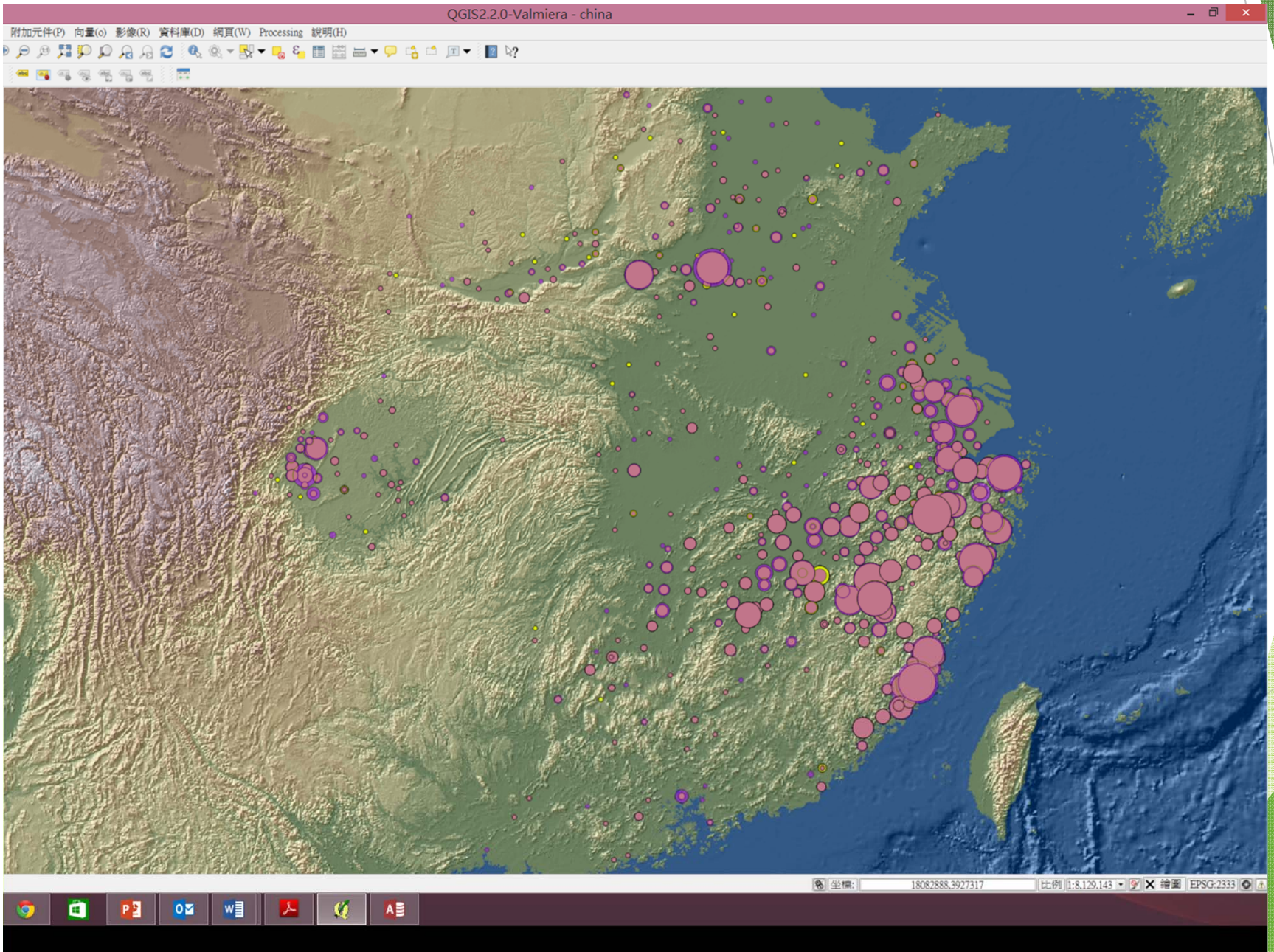
陳淳集中人名社會網絡地理分布

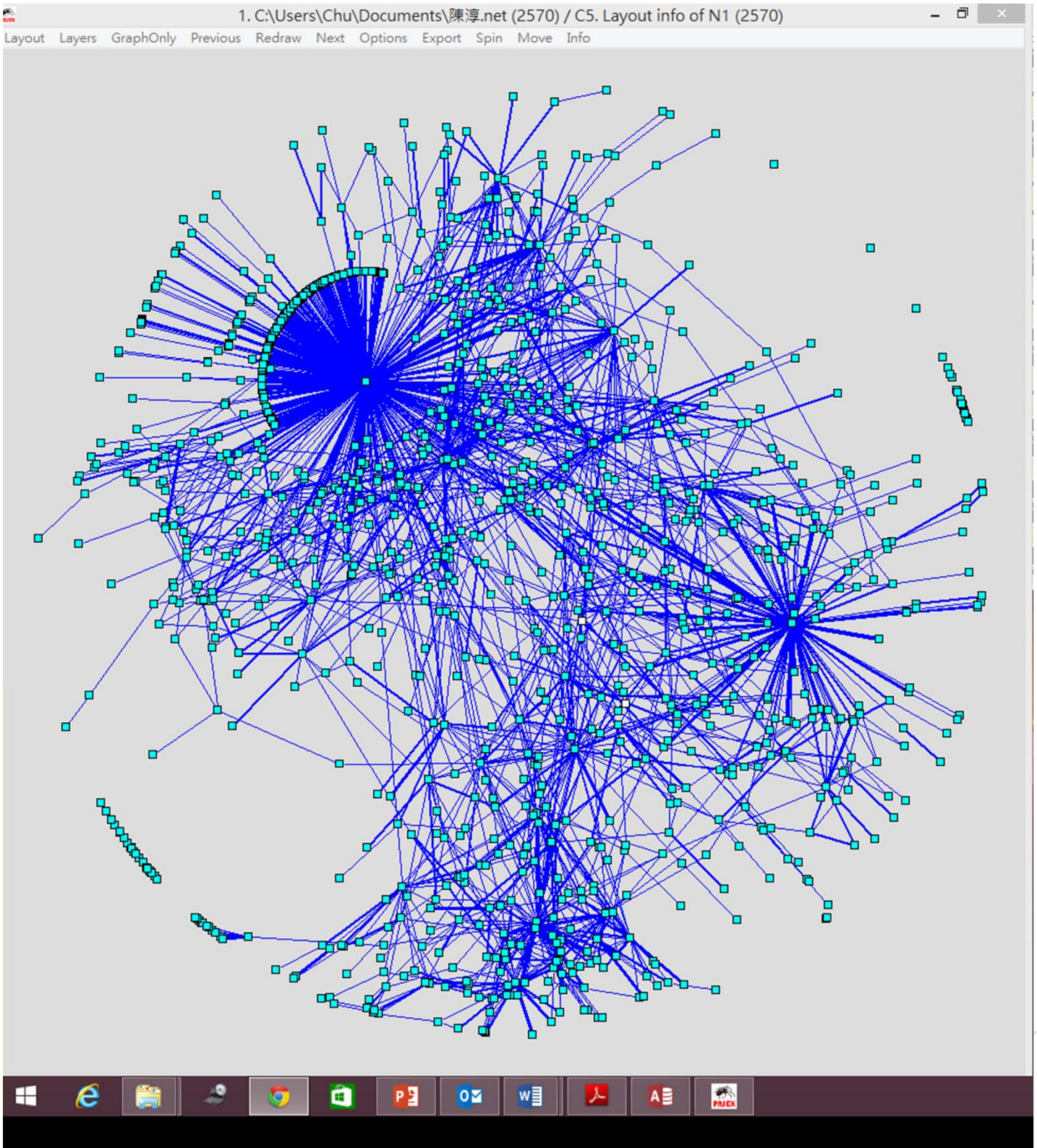


黃幹集中人名社會網絡地理分布

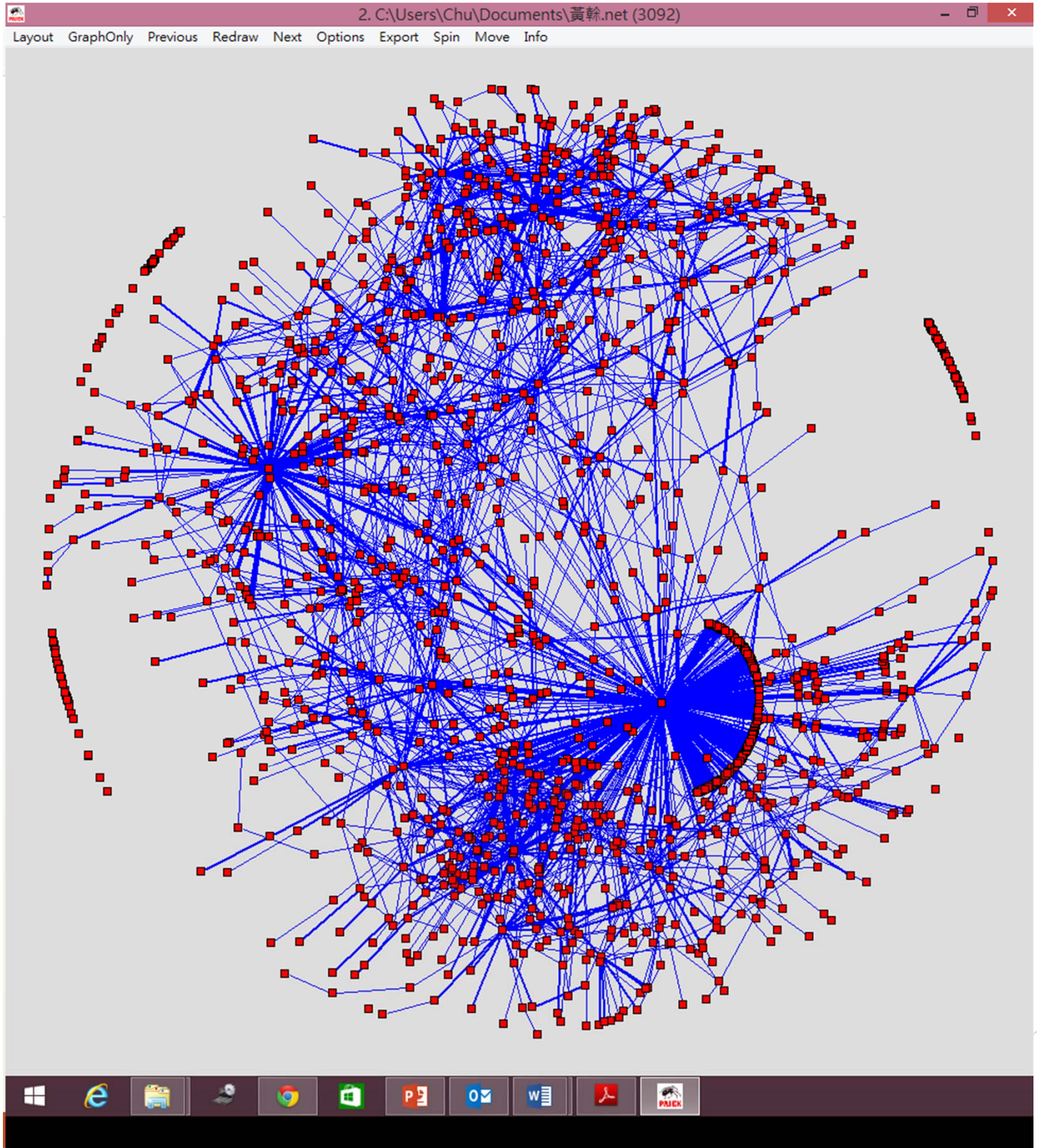


兩人集中人名社會網絡地理分布



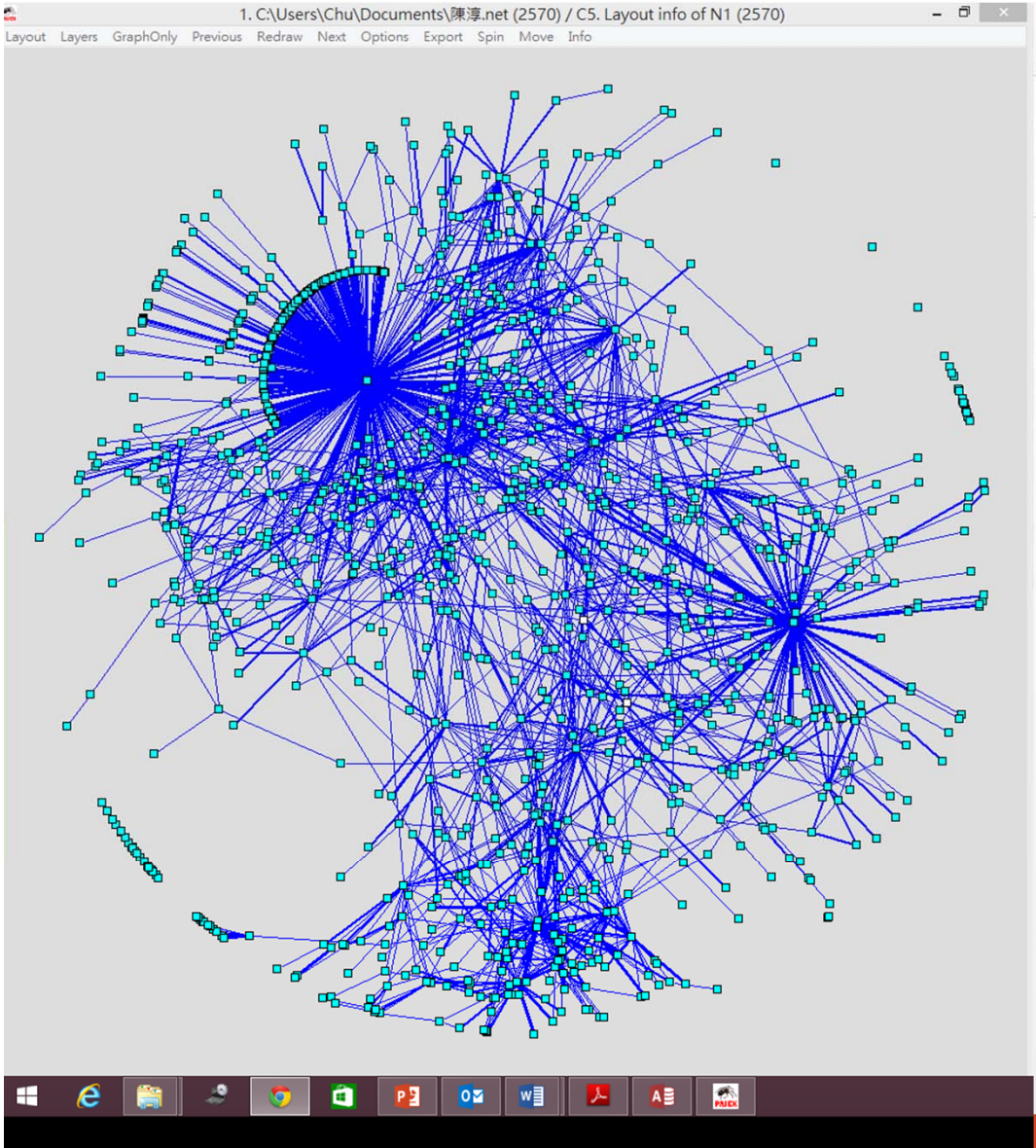


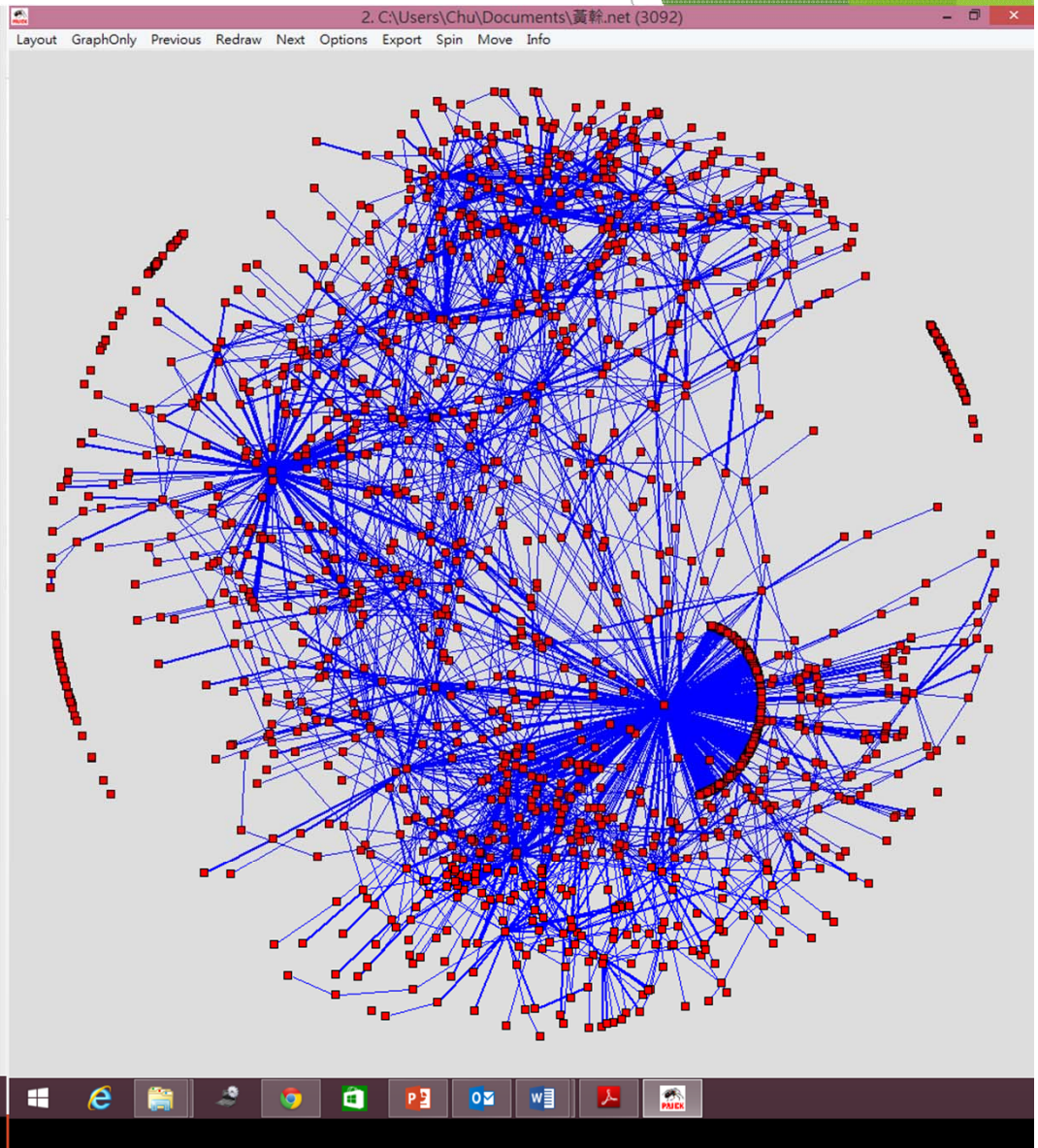
陳淳集中人名學術政治網絡



黃幹集中人名學術政治網絡

比較


1. C:\Users\Chu\Documents\陳淳.net (2570) / C5. Layout info of N1 (2570)
Layout Layers GraphOnly Previous Redraw Next Options Export Spin Move Info


2. C:\Users\Chu\Documents\黃幹.net (3092)
Layout GraphOnly Previous Redraw Next Options Export Spin Move Info

結論

- 一般的全文資料經過XML/TEI標注後可以把原來潛在文本中的資料庫結構以及數據性質明示化。
- 明示化資料庫結構後，可以有各式各樣利用全文資料的方式，亦即它的利用價值會一下子提升很多。
- 但全文資料各式各樣利用的前提是：研究者要能取得全文資料。
  因為不能取得全文資料的話，其利用總是有限的。這樣子也就無法促成對於相同資料進行不同數位解讀的經驗、也難以促成數位人文研究群體的出現。也會使得有全文資料和沒有全文資料變成一種學術不平等。

置入性行銷

- https://sites.google.com/a/ptc.cl.nthu.edu.tw/dhintaiwan/
- http://tinyurl.com/dhintaiwan

終