

# A Cyberinfrastructure for Studying Chinese History

A Proposal Based on the Experience of  
the China Biographical Database  
Project (CBDB)

---

LIK HANG TSUI 徐力恆

China Biographical Database

CUHK DS symposium, 31 March 2017

(Thanks to the help of Peter Bol, Liu Chao-lin, Wang Hongsu)

# Biographies in China

China has over 2000 years of elite biography

There are perhaps more than 250,000 biographies in the historical record: including biographies in dynastic histories, local gazetteers, tomb inscriptions, etc. Local gazetteers (Song-Qing) mention at least 2 million people!

Chinese history is an important source for the history of humanity

So, how do we organize and study this information with modern day technology?



# 中國歷代人物傳記資料庫 (CBDB)

首頁 關於我們▼ 資料來源與涵蓋範圍▼ 方法論▼ 研討會▼ 下載▼ 博客▼ English

歡迎蒞臨中國歷代人物  
傳記資料庫的網站！

#### 簡介

中國歷代人物傳記資料(或稱數據)庫係線上的關係型資料庫，其遠期目標在於系統性地收入中國歷史上所有重要的傳記資料，並將其內容毫無限制地、免

#### 最新通知

有興趣幫我們翻譯**CBDB**的明代官名  
嗎？(2016年11月)

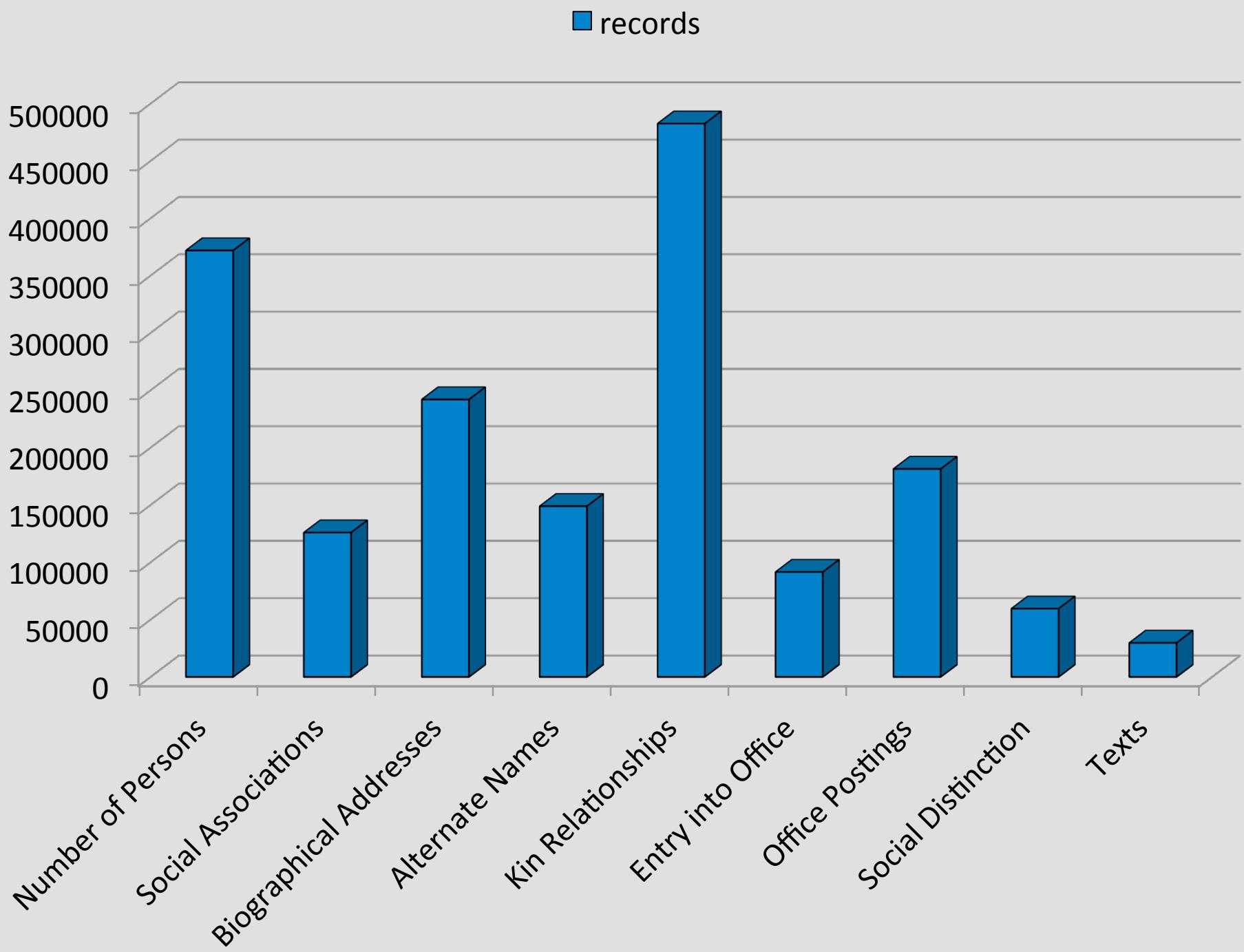
明代官名眾包翻譯系統現已上線！詳情見[此頁](#)。系統將於  
2016年11月7日至2017年6月6日期間運行。

Jointly owned by (since 2005):

Fairbank Center (Harvard), 中國古代史研究中心 (Peking Univ.),  
史語所 (Academia Sinica)

Free and open for academic use

<http://projects.iq.harvard.edu/chinesecbdb>



# Big Data Visualization

## Symposium

9:00 a.m. - 12:35 p.m.

**GIS and Big Data for Urban Applications**

**Prof. HUANG Bo**

Department of Geography and Resource Management  
Faculty of Social Science



# My presentation

- 1. A cyberinfrastructure for Chinese history: what and why?**
- 2. Resource sharing mechanisms**
- 3. Communication between participating actors**

# A Cyberinfrastructure for Historical China Studies (Hongsu Wang, Lik Hang Tsui, Peter K. Bol)

## 服務於中國歷史研究的網絡基礎設施

王宏甦<sup>1</sup>、徐力恒<sup>2</sup>、包弼德<sup>3</sup>

(第七屆數位典藏與數位人文國際研討會論文，2016年12月，臺北)

### 摘要

數據庫、研究項目數量和參與中國文史數位研究的人員大幅增加，使得為中國歷史研究建立相應的網路基礎設施變得必要。網路基礎設施可以起的作用在於連接對一個學科有用的電腦軟件、數據集、人才、實務做法、標準和合作模式，促進研究的進步。本文將具體論述為何要營建中國歷史研究的網路基礎設施，以及如何從資源的共享和成員的交流兩方面實現這個目標。

關鍵詞：網絡基礎設施、中國歷史、數位人文

**A Cyberinfrastructure for Historical China Studies**  
Hongsu Wang, Lik Hang Tsui, Peter K. Bol

**(Paper for the 7<sup>th</sup> International Congress of Digital Archives and Digital Humanities,  
Taipei, Taiwan, Dec. 2016)**

### Abstract

The proliferation of databases for the study of Chinese history and the increasing numbers of researchers taking part in their development calls for a cyberinfrastructure. A cyberinfrastructure can be conceived as a network of discipline-specific software applications and data collections and also of the personnel and the set of best practices, standards, and collaborative methods they establish. This paper discusses how participants in such a cyberinfrastructure for historical China studies can share their resources and how their communication can be facilitated by various technologies and mechanisms.

<sup>1</sup> 哈佛大學「中國歷代人物傳記資料庫」項目經理（Project Manager, CBDB, Harvard University）。電郵為：[hongsuwang@fas.harvard.edu](mailto:hongsuwang@fas.harvard.edu)

<sup>2</sup> 哈佛大學「中國歷代人物傳記資料庫」博士後研究員（Postdoctoral Fellow, CBDB, Harvard University）。電郵為：[tsui01@fas.harvard.edu](mailto:tsui01@fas.harvard.edu)

<sup>3</sup> 哈佛大學副教務長、東亞語言與文明系查理斯·卡威爾（Charles H. Carswell）講座教授（Vice Provost for Advances in Learning and the Charles H. Carswell Professor of East Asian Languages and Civilizations, Harvard University）。電郵為：[peter\\_bol@harvard.edu](mailto:peter_bol@harvard.edu)

**Keywords:** Cyberinfrastructure, Chinese history, digital humanities

### 一、引言：中文數位人文網絡基礎設施的實現

美國學術團體協會（ACLS）在2005年發佈的研究報告《我們的文化共同體》（Our Cultural Commonwealth）提出人文、社會科學應該像自然科學研究一般，有自己的網絡基礎設施（cyberinfrastructure）。<sup>4</sup> 網絡基礎設施的層次介於基礎科技和具體用於某研究項目、某學科和實踐的特定科技之間，可以說處於中層。它可以起的獨特作用，在於連接對一個學科有用的電腦軟件、數據集、人才、實務做法、標準和合作模式。<sup>5</sup> 我們希望在本文處理的議題如下——我們為何要在中國歷史研究領域營建這種網絡基礎設施，以及如何實現這個目標。

和自然科學相比，人文學科和部分社會科學學科（尤其是其中量化特點不明顯的學術領域）深深浸淫在語言之中，很受語言的特點影響。就以主題模型（topic modeling）為例，當這方法用於中國文史研究時會面對頗多挑戰。一般使用這種研究方法時，認定每個詞之間的空格就代表分詞的區隔，但中文文本的情況並非如此。例如，當機器閱讀「中華民國」這四個中文字時，它可以認定那是四個詞，也可以是兩個各為兩個字的詞，也可以是一個四字詞。怎麼讓機器獲得判斷的能力，需要人文學者利用他們的知識介入。

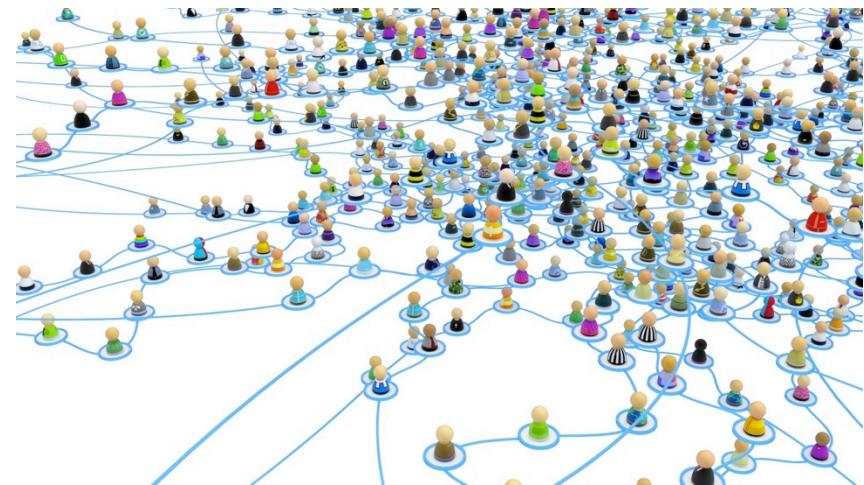
因此，這意味著我們必須深入調查、瞭解語言的特點（例如是古代漢語、現代漢語），才能建立專門用於中國歷史研究的數位人文工具。如果不正視這一點，數位研究計劃之間的溝通和連接將很難進行，數據的分享也會面對很多障礙，導致閉門造車的弊病。不過，近年學界對我們會議的主題——數位人文的興趣越來越濃，用於統計、社會網絡分析、地理空間分析和地圖繪製、文本標註和挖掘、製作主題模型，還有建立關係型數據庫（relational databases）或物件導向式數據庫（object-oriented databases）的電腦軟件如雨後春筍。過去很難用，或者很難獲得的軟件，現在已經是隨手可得，其操作也簡便許多。項目數量和參與數位研究的人員大幅增加，也為中國研究數位人文建立相應的網路基礎設施開始有了成熟的條件。就如其他學科和區域研究的專家一樣，我們領域裡不少學者都感到應該開展這種工作，因此開始了本文的寫作。<sup>6</sup>

<sup>4</sup> 參閱：<http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>。

<sup>5</sup> 參閱：<https://www.nsf.gov/cise/sci/reports/atkins.pdf>。

<sup>6</sup> 比較相關的是佛教研究數據的基礎設施，參閱 Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: The MIT Press, 2015), 186-200.

# What is a cyberinfrastructure?

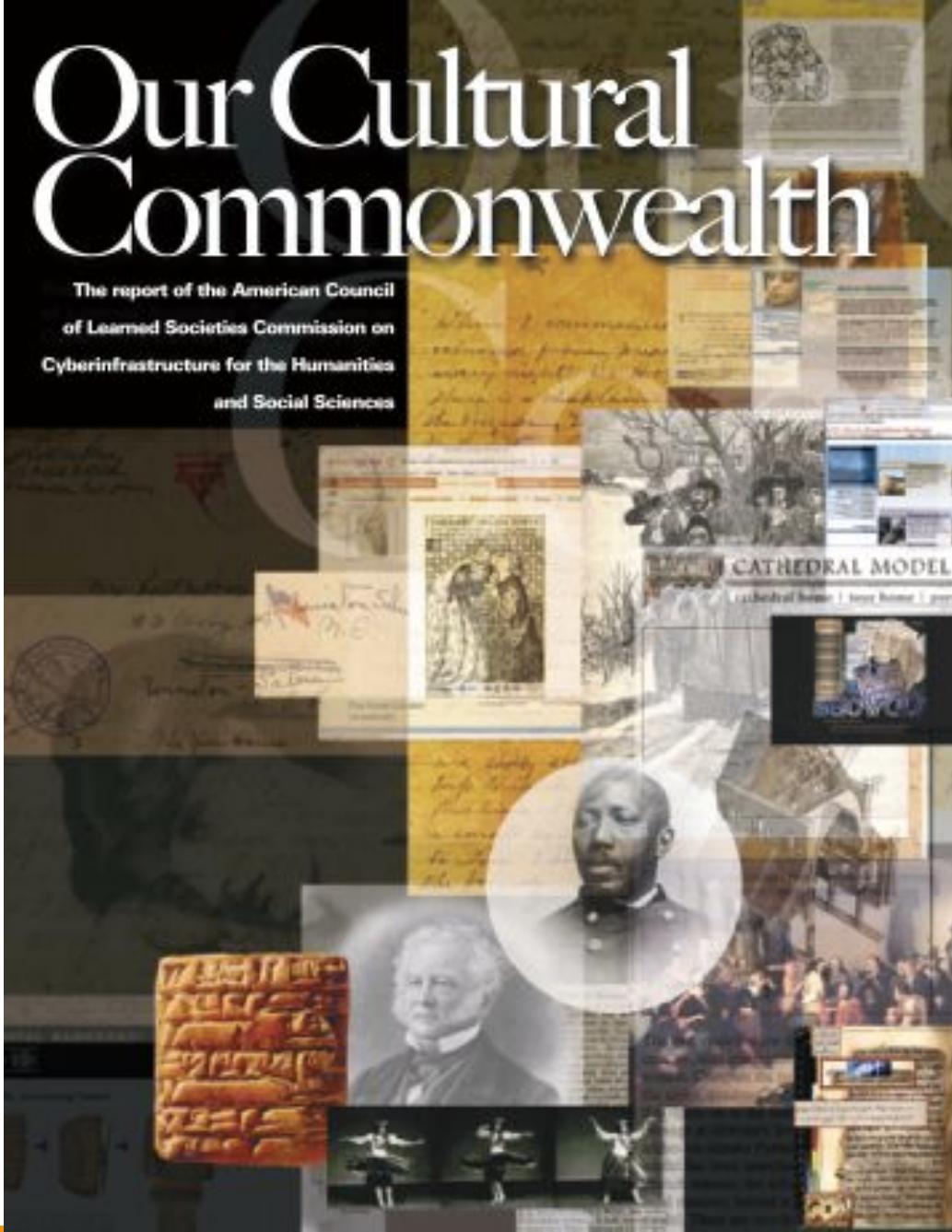


“...research funders use the term **cyberinfrastructure** to describe **research environments** that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services distributed over the Internet **beyond the scope of a single institution.**”

(<https://en.wikipedia.org/wiki/Cyberinfrastructure>)

# Our Cultural Commonwealth

The report of the American Council  
of Learned Societies Commission on  
Cyberinfrastructure for the Humanities  
and Social Sciences



## ACLS report (2005)

- Tangible:

- network and means of storage in digitized form, discipline-specific software applications and project-specific data collections.

- Intangible:

- layer of expertise and the best practices, standards, tools, collections and collaborative environments that can be broadly shared across communities of inquiry.

盡平上思州黃勝許與交趾表裏寇邊二十九年國  
桀率兵討之賊走象山山近交趾深林不可入乃度  
其出入列柵圍之徐伐山通道且戰且進二年拔其度  
寒榜許走交趾國傑以巢地爲屯田募度遠諸種人  
耕之以爲兩江蔽障

**蕭泰登**字則平吉安路人至元三十一年年  
論死者二百人泰登錄之而釋不知情者百三十  
七人他所辨雪無算凡黜貪謬者二百一十人

**謝讓**字仲和潁昌人大德中爲湖廣行省左右司郎中  
時廣西兩江岑雄黃勝許等屢相讐殺爲邊患讓  
謂此曹第可懷柔不宜力競寬其法以羈縻也

**尼**字尚文唐兀氏人至正末爲廣西行省平章政事  
兼廉訪使時紅巾擾攘吉尼以公費修築城池民事  
不知勞洪武元年楊璟取廣西吉尼堅壁不下

**陳瑜**字仲庸雷州人廣西中書省都事城破以佩刀自刎  
城破執送京師不屈死郡人感其德立廟祀之

**劉永錫**字潭州人與瑜同事率妻子溺於白龍池死焉

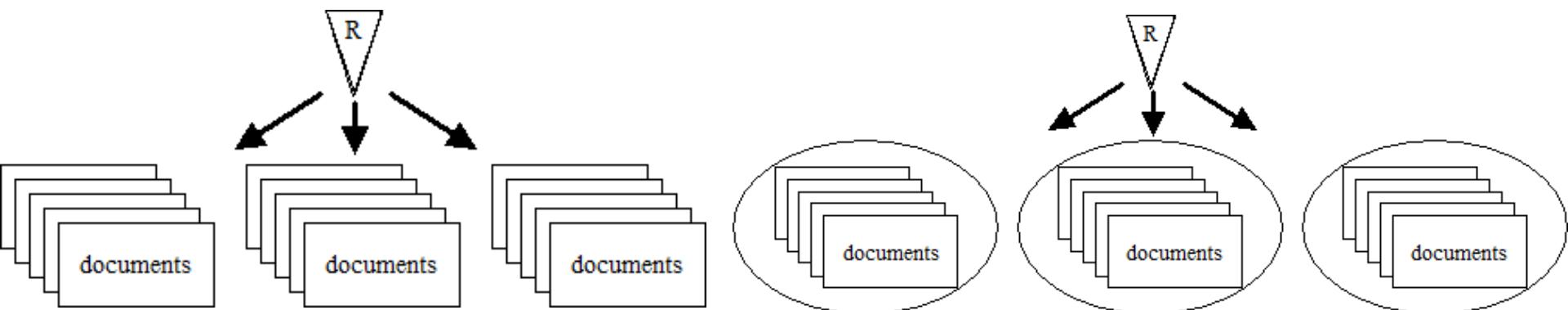
**曾尙賓**江西人爲義兵千戶洪武元年明兵圍靜  
江尙賓守西城城陷身中數鎗知不敵自刎

# Traditionally...

---

Researchers gather their own research material by themselves, through their own ways.

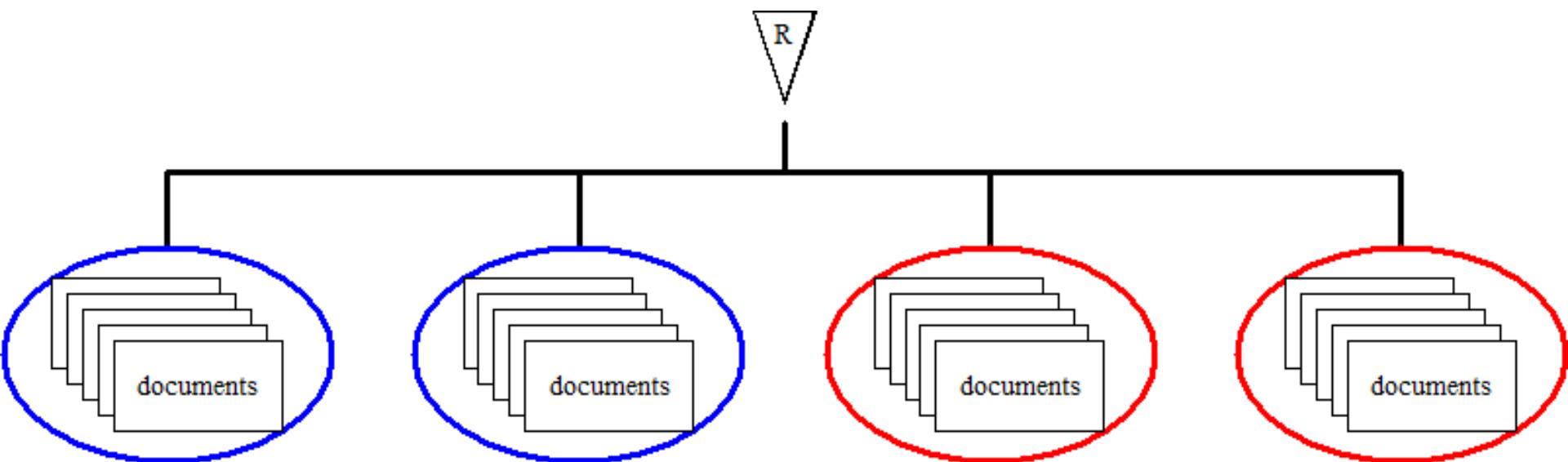
Sometimes they also relied on collections in libraries.



# Then, with the Internet and the digital age...

---

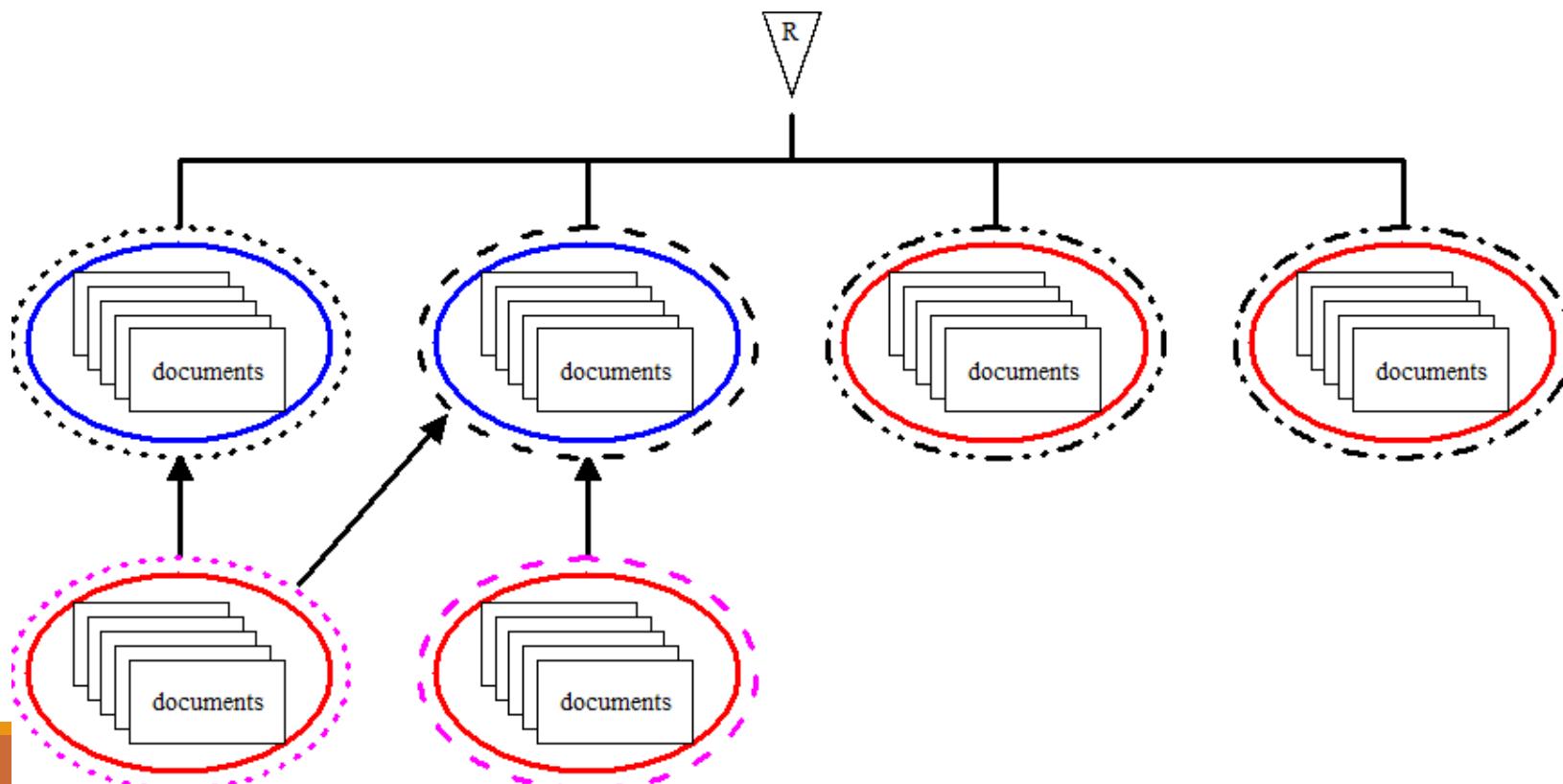
Some materials are open, some are commercial, some are restricted



# And usually more like this:

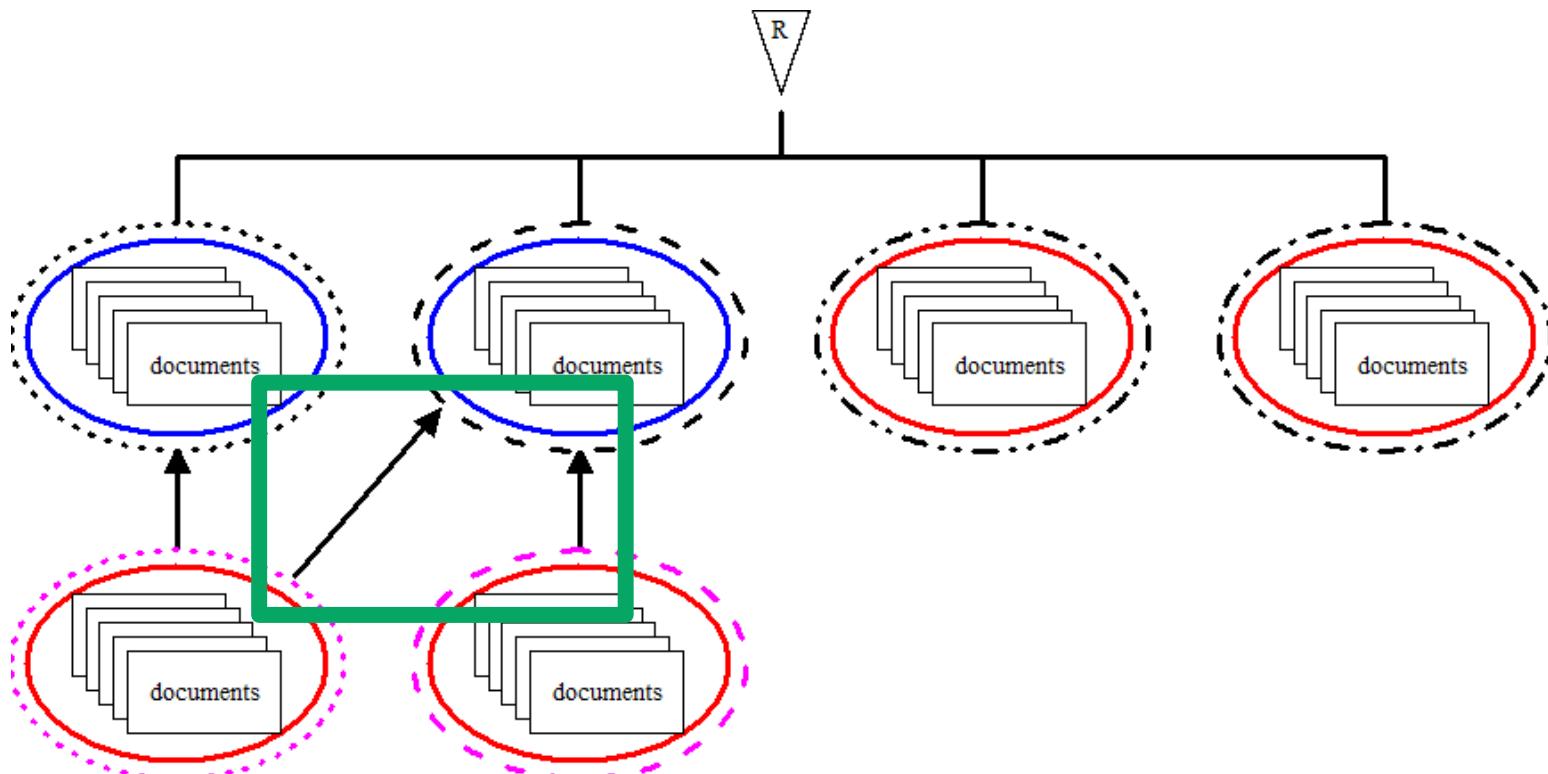
---

Different collections, different formats, different levels of access, different interfaces for using research data



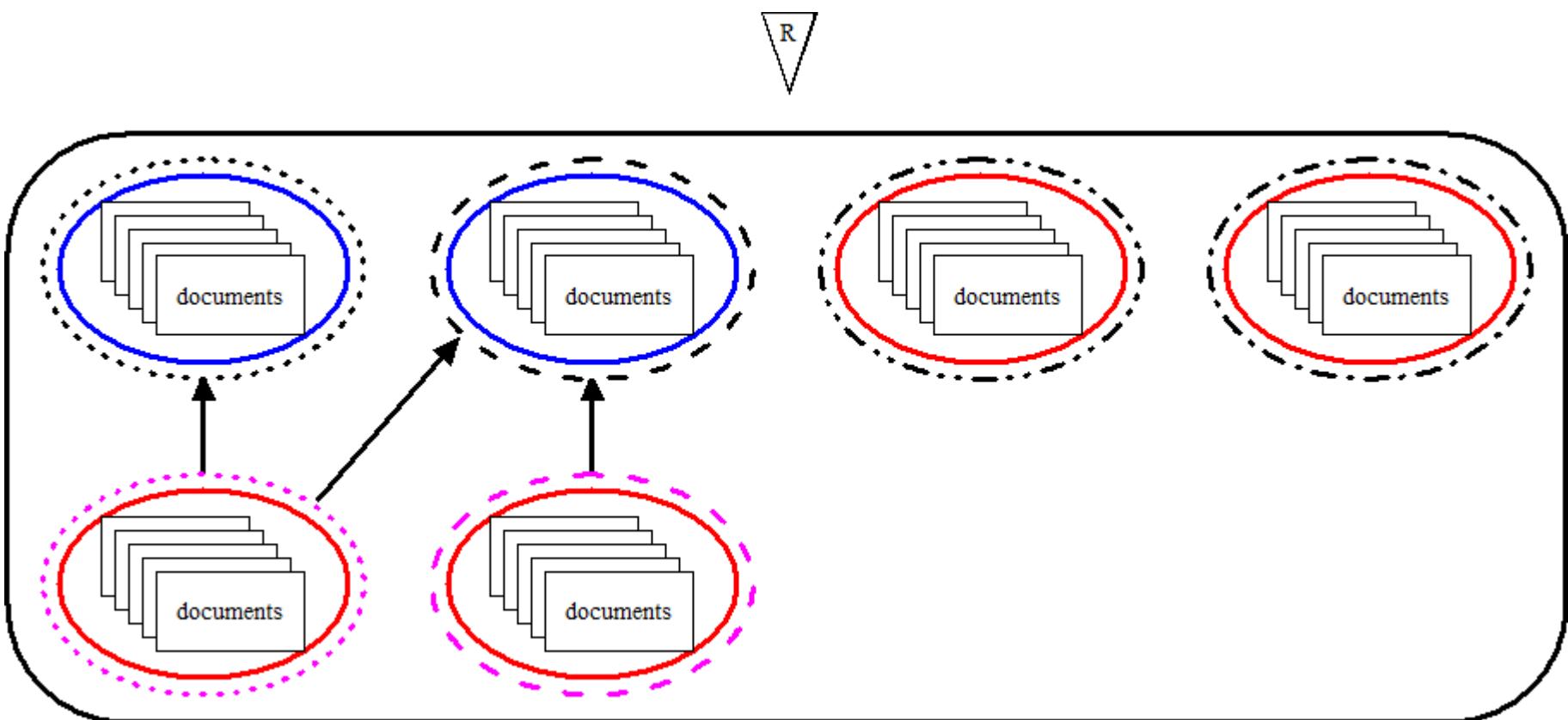
# Libraries help us integrate some of these interfaces

---



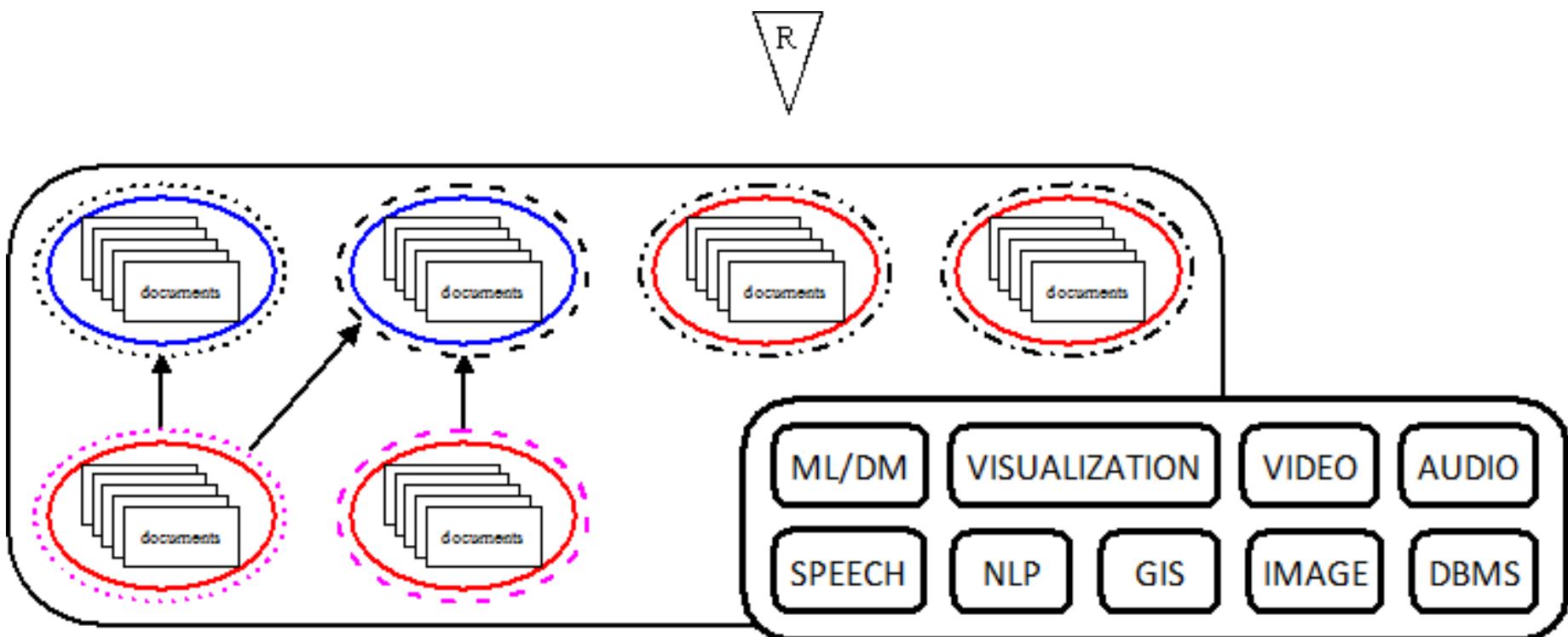
# Ideally, a cyberinfrastructure helps us access different data sources in compatible ways

---



# And also provide tools for using these data

---



So, how?

**Resource sharing + communication**

## Resource sharing:

1. Sharing through APIs
2. The sharing of documents

# APIs: Application Programming Interface

---

APIs allow users access resources through a relatively stable program interface.

When the providers modify, enhance, or expand the resources, the APIs generally remain the same, thereby relieving users of unnecessary technical involvements.

Users specify the formats of their requests via an input interface, and obtain the requested resources via an output interface.

The syntax and semantics of the interfaces must be defined precisely.

```

- <Package>
- <PersonAuthority DataSource="CBDB">
- <PersonInfo>
- <Person>
- <BasicInfo>
    <PersonId>1762</PersonId>
    <EngName>Wang Anshi</EngName>
    <ChName>王安石</ChName>
    <IndexYear>1080</IndexYear>
    <Gender>0</Gender>
    <YearBirth>1021</YearBirth>
    <DynastyBirth>北宋</DynastyBirth>
    <EraBirth>元祐</EraBirth>
    <EraYearBirth>5</EraYearBirth>
    <YearDeath>1086</YearDeath>
    <DynastyDeath>北宋</DynastyDeath>
    <EraDeath>元祐</EraDeath>
    <EraYearDeath>1</EraYearDeath>
    <YearsLived>66</YearsLived>
    <Dynasty>宋</Dynasty>
    <JunWang>太原</JunWang>
    <Source>宋人傳記資料索引(電子版)</Source>
    <Pages>1536</Pages>
+ <Notes></Notes>
</BasicInfo>
- <PersonAliases>
- <Alias>
    <AliasType>字</AliasType>
    <AliasName>介甫</AliasName>
</Alias>
- <Alias>
    <AliasType>室名、別號</AliasType>
    <AliasName>半山老人</AliasName>
</Alias>
- <Alias>
    <AliasType>諡號</AliasType>
    <AliasName>文</AliasName>
</Alias>

```

```

    "Package" : {
        "PersonAuthority" : {
            "DataSource" : "CBDB",
        },
        "PersonInfo" : {
            "Person" : {
                "BasicInfo" : {
                    "PersonId" : "1762",
                    "EngName" : "Wang Anshi",
                    "ChName" : "王安石",
                    "IndexYear" : "1080",
                    "Gender" : "0",
                    "YearBirth" : "1021",
                    "DynastyBirth" : "北宋",
                    "EraBirth" : "元祐",
                    "EraYearBirth" : "5",
                    "YearDeath" : "1086",
                    "DynastyDeath" : "北宋",
                    "EraDeath" : "元祐",
                    "EraYearDeath" : "1",
                    "YearsLived" : "66",
                    "Dynasty" : "宋",
                    "JunWang" : "太原",
                    "Source" : "宋人傳記資料索引(電子版)",
                    "Pages" : "1536",
                    "Notes" : "Wang(2) Anshi [1762] Yi(3)'s [7082] son, Gu
                },
                "PersonAliases" : { "Alias" : [
                    {
                        "AliasType" : "字",
                        "AliasName" : "介甫"
                    },
                    {
                        "AliasType" : "室名、別號",
                        "AliasName" : "半山老人"
                    },
                    {
                        "AliasType" : "諡號",
                        "AliasName" : "文"
                    },
                    {
                        "AliasType" : "小字",
                        "AliasName" : "孺郎"
                    }
                ] },
                "PersonAddresses" : [

```

E.g. CBDB's API

# CTEXT

百諸家子 Chinese Text Project

Post-Han -> Sui-Tang -> 道典 -> 選舉二

《選舉二》

《歷代制中》

1 [x] 歷代制中: 魏 晉 東晉 宋 齊 梁 陳 後魏 北齊 後周 隋

2 [x] 歷代制中: 魏文帝為魏王時，三方鼎立，士流播遷，四人錯雜，詳覈無所。延康元年，吏部尚書陳群以天朝選用不盡人才，乃立「九品官人之法」，州郡皆置中正，以定其選，擇州郡之賢有識鑒者為之，區別人物，第其高下。又制：郡口十萬以上，歲察一人，其有秀異，不拘戶口。初，曹公時，魏府初建，以毛玠、崔燭為東都錄史，銳衡人物，選用先後勤儉。於是天下士人皆砥礪名節，務從勸懲。和洽言於公曰：「天下大器，在位與人，不可以一節檢也。俊素過中，自以處身則榮，以此格物，所失或多。今朝廷之儀，必有著新衣、乘好車者，不謂之廉潔。至令士大夫故污辱其衣，蔽其與服，朝府大吏或自挈壘祿，以入官署。夫立教觀俗，貴處中庸，為可繼也。今崇一概難堪之行，以檢殊途，勉而為之，必有變疾。古之大教，務在通人情而已。凡激節之行，則容僞矣。」其武官之選，俾護軍主之。黃初三年，始除舊漢限年之制，令郡國貢舉，勿拘老幼，儒通經術，吏達文法，到皆試用。

3 [x] 歷代制中: 自明帝太和之後，俗用浮靡，遞相標目，而夏侯、諸葛、何、鄧之儔，有四聰八達之稱，帝深所嫉之。於是，惡士大夫之有名聲者，或禁錮廢黜以微之。吏部尚書盧毓奏曰：「古者敷奏以言，明試以功。今考績之法久廢，而毀稱相進退，故真偽混雜也。」帝遂詔散騎常侍劉劭作都官考課之法，以考覈百官。具考績篇。

4 [x] 歷代制中: 齊王嘉平初，曹爽既誅，司馬宣王秉政，詳求理本。中護軍夏侯玄言曰：「夫官才用人，國之柄也。故銳衡專於臺閣，上之分也；孝行考乎閭巷，優劣任之鄉人，下之敘也。夫欲清教審選，在明其分敘，不使相涉而已。今令中正但考行倫輩，輩當行均，斯可官矣。行有大小，比有高下，則所任之次亦然別矣。奚必使中正干銳衡之職於下，而執機柄者有所委仗於上，上下交侵，以生紛錯哉？且眾職之屬，各有官長，但使官長各以其屬能否賦之；臺閣則據官能否之定，參以鄉閭德行之次，擬其倫比，勿使偏頗；中正則唯考行跡，別其高下，審定輩類，勿使升降，而總之於臺閣，官長所第，中正輩擬，比隨次率而用之。如其不稱，責在負外。則內外相參，得失有所，庶可靜風俗而審官才矣。」兼請除重設之官，定服制之等。宣王辭不能改，請俟於他質。按，九品之制，初因後漢建安中天下兵興，衣冠士族多離本土，欲徵源流，慮難委悉，魏氏革命，州郡縣俱置大小中正，各取本處人任諸府公卿及臺省郎吏有德充才者為之，區別所管人物，定為九品。其有言行修整，則升進之，或以升五品，以六升五；儻或道善虧體，則降下之，或自五退六，自六退七矣。是以吏部不能審定要。

# CBDB

Welcome to China Biographical Database (CBDB)

CHINA BIOGRAPHICAL DATABASE PROJECT (CBDB)  
中國歷代人物傳記資料庫

Look at Entry

NAVIGATION\_PANE

All Access O... < > File Home Create External Data Database Tools Tell me what you want to do

Library Resources Library Resources

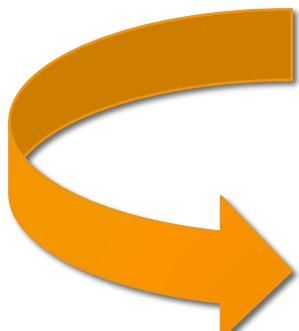
Search... Tables ADD\_CODEs ALNAME\_CODES APPOINTMENT\_TYPE... ASSOC\_CODE\_TYPE... ASSOC\_CODES ASSOC\_CODES ADDRESS\_OFICE\_C... BIOG\_ADDR\_CODES BIOG\_INST\_CODES BIOG\_MAIN CHRONICAL\_CODES CopyTables COUNTRY\_CODES DYNASTIES ENTRY\_CODE\_TYPE... ENTRY\_CODES ENTR\_TYPES ETHNICITY\_TRIBE\_C... EVENT\_CODES EXANTIC\_CODES FormLabels GANZH\_CODES HOUSEHOLD\_STAT... KIN\_Mourning KIN\_MOURNING\_ST... KINSHIP\_CODES LITERARYGENRE\_CO... MEASURE\_CODES NameAutoCorrect ...

Select Entry Examination: jinshi (general) 科舉: 進士(舉人) From: 1000 Use Dates To: 1100 Use Index Years [[Filter]] Import Places All Places Use XY Reference

Type: N/A

Name	姓名	Index Year	Entry	入仕年	From	官位	地址類別
Zhu Shizhen	朱士珍	1093	1065 examination: jinshi (general)	科舉: 進士(舉人)	Xiping	新平	籍貫(本地地級)
Lu Daizhong	呂大忠	1089	1053 examination: jinshi (general)	科舉: 進士(舉人)	Lantian	蘭田	籍貫(本地地級)
Sima Chi	司馬芝	1039	1050 examination: jinshi (general)	科舉: 進士(舉人)	Xia Xian	夏縣	籍貫(本地地級)
Sima Guang	司馬光	1078	1038 examination: jinshi (general)	科舉: 進士(舉人)	Xia Xian	夏縣	籍貫(本地地級)
You Shizhong	游酢	1097	1065 examination: jinshi (general)	科舉: 進士(舉人)	Wugong	武功	籍貫(本地地級)
Zhang Sheng	張昇	1051	1015 examination: jinshi (general)	科舉: 進士(舉人)	Hancheng	韓城	籍貫(本地地級)
Zhang Jufu	張居	1121	1091 examination: jinshi (general)	科舉: 進士(舉人)	Zhening	真寧	籍貫(本地地級)
Li Fu	李復	1109	1079 examination: jinshi (general)	科舉: 進士(舉人)	Changan	長安	籍貫(本地地級)
Guan Shi	關思	1063	1030 examination: jinshi (general)	科舉: 進士(舉人)	Changan	長安	籍貫(本地地級)
Wang Zheng	王鑒	1028	1019 examination: jinshi (general)	科舉: 進士(舉人)	Xiangyang	襄陽	籍貫(本地地級)
Bai Yue	白約	1083	1053 examination: jinshi (general)	科舉: 進士(舉人)	Rong Zhou	榮州	籍貫(本地地級)

Run Query Save to GIS GR818030 UTF-8 Display Language: 繁體 簡體 Help Exit



# MARKUS

MARKUS Save to system Back to last save Export Tools Links FAQ HowTo Login

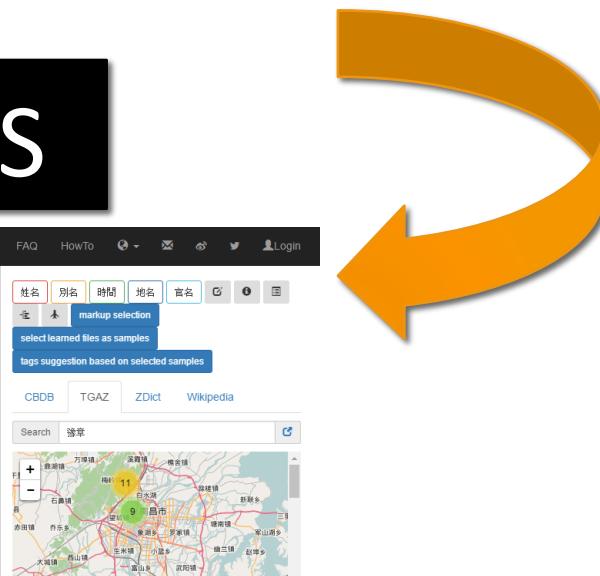
人，如于湖、石湖，止齋者。亦有能詩時賦、時論、記跋之類者，往往耽商工。言禱福，卻多不驗。

近時都下有土人，許其姓者，能迎祀大仙，所言多奇。

【嘉定壬午】之春，三山董公朴、同一、二朋友許，扣功名大駕。即書「沙門光遠降」許。先作自贊云：「備脚自由，屈腳自在。不知十二部尊經，不識三千條大戒。醉吟高歌，無彈無擣。當時若見祖師，任它祖師扭頭。」又云：「無疑無疑，自有東西。目前有險，眼下阿鼻。不認真實法體，不念如意菩薩。捉金毛獅子，任願免免如飴。」許「徐公書云：「董子！董子！文魁博士，醜醜胸中十萬兵。無垢為朋，汪（公相）似。若得火土相逢，一躍萬頭浪濶。」許「後人」，歲在癸丑，董公赴大魁天下。董生於壬子，魁乙丑，審得火土相克之句，就元局中，諱公唱和時，有醜積、縱橫一聯，不差一字，黃生已先知之矣。光遠万昔時曾遊入閩，至青城山丈人觀，不為道士所禮。偽為破衣出巡，盡醉觀中向來不禮之人，南僧不計入閩者以此。距今百餘載，尚為黠兔，可謂異事。

【龍溪】先生汪公藻，字彦章，吾郡之德興人。幼年已負文名。作詩云：「一春踏蕪千日耕，處處耘雲將雨行。野田青水碧於鏡，人影波倒晴不驚。桃花偶然出蕊苔，似開未開最有情。茅茨烟曉杏衣濃，破夢牛鶴啼一聲。」汪公詩第一出，傳為詩社公所稱。時年宰落，莫究所學。朱叔止題其墨云：「名高從昔號龍溪，未免羞兒著力讚。」一日，鶯心倦極，一日，雨霽深隱愚溪。不逢誰旦闢昌運，終抱沈埋故故。已矣九原寧可作，蕭蕭古木亂蟬嘶。朱叔止題」亦為詩公所稱。抑止名紙，舍人仲之註。

【水福禪】之東南八十里，瀘漢寺之巖，有茶樹十。形體奇怪，環布巖石。不著姓名，人所未知。號曰「仙茶」。歐陽公水叔嘗得之，喜其無蕊結之跡，如指畫成文。欲以書字畫號譯之，未暇也。蔡（滿清）時守三山，以道家書詣其曰：「首道守真一，中有不朽死術。」蔡（滿清）亦莫得其據。政和三年之夏，邑宰陳祐，好奇之士也。訪求其詳，知篆有三：一在安仁寺仙人山，寺僧僅蠶蠶之，燭蠶而庵之；二在中和寺舊坑之巖，今存數株，字皆奇絕。而不知其名。三在瀘漢之山巖也。安仁者，掘而得之，僅完二字，又於上生篆留範，得所藏篆書之餘，復再續篆體，列於巖上。今閱之，余得其意。余嘗見碑文，字勢夭矯，瀟洒奇妙。枝葉不墨，而解綵皆通，信是奇怪。不知忠惠題道家何書而識之？此字恐子雲未必識也。



hvd.32581  
豫章郡 (Yuzhang Jun)  
(201 ~ 13) [115.89772, 28.67490]  
  
hvd.97186  
豫章郡 (Yuzhang Jun)  
(201 ~ -125) [0.00000, 28.67490]

# China Historical GIS (TGAZ) API

## 中國歷史地理信息系統API

### TGAZ API

[intro](#)

[usage](#)

[examples](#)

[credits](#)

[cga](#)

:: home

Chinese historical records the valid years of the database are -222 to 1911.

## Canonical Placename

- [http://maps.cga.harvard.edu/tgaz/placename/hvd\\_32180](http://maps.cga.harvard.edu/tgaz/placename/hvd_32180)

### Canonical Placename Formats

- **json** [http://maps.cga.harvard.edu/tgaz/placename/json/hvd\\_80547](http://maps.cga.harvard.edu/tgaz/placename/json/hvd_80547)
- **rdf** [http://maps.cga.harvard.edu/tgaz/placename/rdf/hvd\\_135744](http://maps.cga.harvard.edu/tgaz/placename/rdf/hvd_135744)
- **html** [http://maps.cga.harvard.edu/tgaz/placename/html/hvd\\_9732](http://maps.cga.harvard.edu/tgaz/placename/html/hvd_9732)
- **xml** [http://maps.cga.harvard.edu/tgaz/placename/xml/hvd\\_96066](http://maps.cga.harvard.edu/tgaz/placename/xml/hvd_96066)

## Faceted Search

parameters allowed:

- **n**: name (the spelling of the placename)
- **yr**: year of existence (the year during which the placename existed)
- **ftyp**: feature type (the feature type, or class of placename)
- **src**: source (the data source, such as CHGIS, RAS)
- **p**: part of (the immediate parent jurisdiction where the place was located)
- **fmt**: format (the output format returned: xml, json, html)

### Placename Examples

- Pinyin: <http://maps.cga.harvard.edu/tgaz/placename?n=Tianbian>
- Chinese: <http://maps.cga.harvard.edu/tgaz/placename?n=晋阳>
- Tibetan: <http://maps.cga.harvard.edu/tgaz/placename?n=%E9%9D%99>
- Russian: <http://maps.cga.harvard.edu/tgaz/placename?n=Вятское Наместничество>

# Sharing (or even crowdsourcing) place name data?

CHGIS

(Harvard CGA+  
Fudan 史地所)

Base maps,  
times series  
data for place  
names.

**China Historical GIS**



Gazetteer Search Engine

List of Free Datasets

Skinner Map Collection

with funding from the Henry Luce Foundation, the National Endowment for the Humanities and support from

Fairbank Center for Chinese Studies  
Harvard Asia Center  
Harvard Yenching Library  
Harvard Yenching Institute

© 2001-2010 - Harvard University and Fudan University

CHGIS  
Center for Geographic Analysis  
1737 Cambridge St, K021  
Cambridge, MA 02138  
tel: 617-496-9439  
Geohash36: 9LM3XHCKLR  
office location

China GIS DATA

Japan GIS Demo

Dynamic Maps WEB MAPS

Map Scans 地圖

Featured EdSite

Sichuan Earthquake 8級地震

The "List of Free Datasets" button is highlighted with a red oval.

<http://www.fas.harvard.edu/~chgis/>

# APIs and Data Crawling

---

- To enable big data methods
  - e.g. distant reading
- Crawling/mining data  $\neq$  stealing data
- More complex needs in terms of data:
  - not just full-text searches
- Agreements between projects to establish data crawling rights

# Moretti: “遠讀”

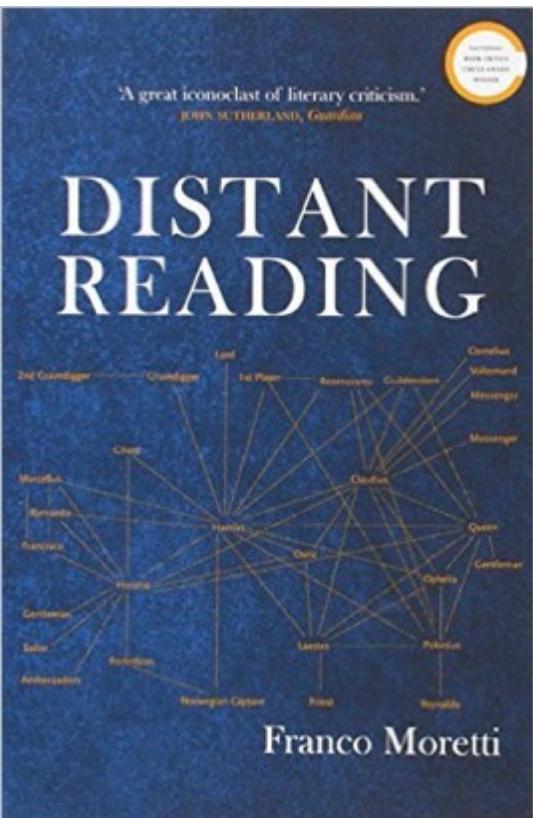


FIGURE 1: *The rise of the novel, 18th to 20th century*



New novels per year, by 5-year average. Sources: For Britain: W. H. McBurney, *A Check List of English Prose Fiction, 1700–99*, Cambridge, MA 1960, and J. C. Beasley, *The Novels of the 1740s*, Athens, GA 1982; both partly revised by James Raven, *British Fiction 1750–70: A Chronological Check-List of Prose Fiction Printed in Britain and Ireland*, London 1987. For Japan: Jonathan Zwicker, 'Il lungo Ottocento del romanzo giapponese', in *Il romanzo*, vol. III, *Storia e geografia*, Torino 2002. For Italy: Giovanni Ragone, 'Italia 1845–70', in *Il romanzo*, vol. III. For Spain: Elisa Martí-López and Mario Santana, 'Spagna 1845–1900', *Il romanzo*, vol. III. For Nigeria: Wendy Griswold, 'Nigeria 1950–2000', *Il romanzo*, vol. III.

To enrich our literary chronicles with a few new historical ingredients . . . would be pointless: it's the presuppositions which must change, and the object transform itself. To abolish the individual from literature! It's a laceration, clearly, even a paradox. But a literary history is possible only at this price.

Roland Barthes, 'History or Literature?'

Home    About    Latest    Our books    Series    Resources    LSE Comment    Popular    

## The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK.



New copyright exceptions to text and data mining for non-commercial research have recently come into effect and this is welcome news for UK researchers and research, argues **Ross Mounce**. Here he provides a brief overview of the past issues discouraging text and data mining and the what the future holds now that these exceptions have been introduced. But despite legal barriers being removed, many technical barriers still remain. Furthermore it remains to be decided what formally constitutes 'non-commercial' research.

After eight long years including not one but two expert-led reviews of intellectual property; new **copyright exceptions**, some of which in particular will enable and empower UK academic research



Email Address

Subscribe to the Impact Blog



This work is licensed under a  
**Creative Commons Attribution  
3.0 Unported License** unless  
otherwise stated.

# DIGGING INTO DATA CHALLENGE

Mining  
biographies in  
2,000 local  
gazetteers  
地方志

## Automating Data Extraction from Chinese Texts

### PROJECT

about  
team  
links

### ANALYSIS

tools  
[data](#)

### EVENTS

workshops  
release

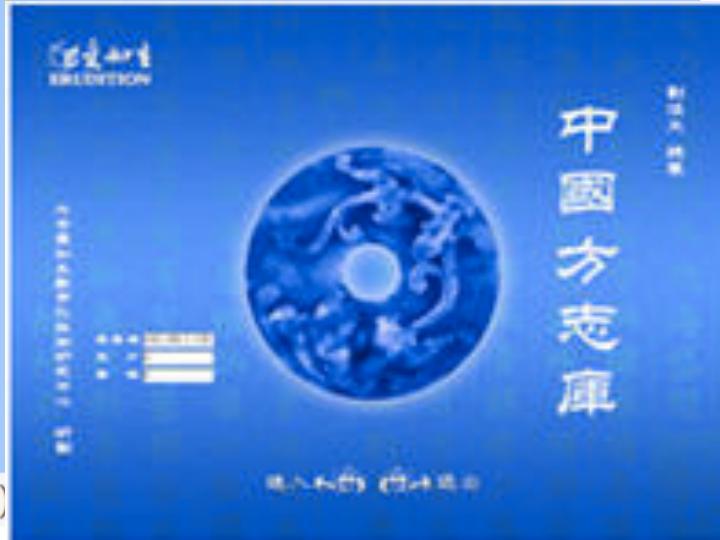
Structured datasets extracted from 41 early gazetteers held by the Institute of History and Philology at Academia Sinica and 2,000 later gazetteers will be made available here.

### Gazetteers

Local gazetteers are massive repositories of data pertaining to mountains and rivers, roads and bridges, administrative divisions, taxes and population, agricultural production, disasters and epidemics, notable local people, government officials, public and private schools, tombs and monuments, religious sites, selected writings of local authors, and a host of other topics. They were first introduced at the prefectural level (about 350 prefectures), were slowly adopted by counties (about 1300), and eventually were even written for some towns and religious sites. Authors of gazetteers focused primarily on the amassing of data rather than interpretation of events, and the categories of information they collected remained fairly consistent since the twelfth century. Individual researchers have used gazetteers to gather information on particular topics, such as natural disasters, examination candidates, and religious sites. However, systematic, national-level studies based on gazetteer data are very rare because the process of manually collecting data is extremely laborious, cannot be easily altered once begun, and is difficult to replicate in related studies.

### Recent Tweets

-  @Tom Mullaney  
 @DID\_ACTE  
Now Live #DHAsia @cesta\_stanford A Macroanalysis of Modern Japanese and Chinese Texts @RichardJeanSe @hoytlong on <https://t.co/Sx8WePJYXV> 3 weeks, 1 day ago
-  @DID\_ACTE  
Chinese text analysis and reading platform MARKUS featured on H-ASIA's Digital Asia resources @HNet\_Humanities <https://t.co/N95Ts16dI> #DH 3 weeks, 3 days ago
-  @DID\_ACTE  
Also featuring Chinese humanities & social sciences, and #MARKUS ! <https://t.co/1CifZPqTR> 3 weeks, 4 days ago



Book1	Book2	overlap .count	count.adj .1	count.adj .2	percentage overlap	percentage overlap		Y: overlap N:separate
平湖縣志	平湖縣志	266	307	2688	86.64495114	9.895833333	86.64495114	Y
元城縣志	大名縣志	422	496	1766	85.08064516	23.89580974	85.08064516	Y
宣城縣志	宣城縣志	462	583	550	79.24528302	84	84	Y
滄州志	重修天津府志	506	605	7037	83.63636364	7.190564161	83.63636364	Y
烏程縣志	烏程縣志	437	688	546	63.51744186	80.03663004	80.03663004	Y
固始縣志	固始縣志	203	254	449	79.92125984	45.21158129	79.92125984	Y
雷州府志	徐聞縣志	297	2019	432	14.7102526	68.75	68.75	Y
重修天津府志	南皮縣志	357	7037	534	5.073184596	66.85393258	66.85393258	Y
溫州府志	平陽縣志	485	3578	811	13.55505869	59.8027127	59.8027127	Y
義烏縣志	金華府志	450	866	1771	51.9630485	25.40937324	51.9630485	Y
海豐縣志	陸豐縣志	103	632	302	16.29746835	34.10596026	34.10596026	Y
長泰縣志	八閩通志	133	483	14762	27.53623188	0.900961929	27.53623188	Y
八閩通志	漳平縣志	24	14762	90	0.162579596	26.66666667	26.66666667	Y
八閩通志	內黃縣志	28	14762	132	0.189676196	21.21212121	21.21212121	N

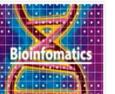
# Sharing of Documents

---

- Sustainability of data:
  - standards for long term storage e.g. Dataverse, ICPSR
- Version control:
  - not fixed entities
  - encouraging citation of these resources
- Description:
  - user's guide, metadata descriptions etc.

 北京大学 开放研究数据平台  
Peking University Open Research Data

Q Featured Dataverses

 China Survey Data Archive	 China Family Panel Studies	 China Health and Retirement Longitudinal Study, CHARLS	 The Research Center For Contemporary China
 Center for Healthy Aging and Development Studies	 visualization and visual analytics research group, Peking University	 Center for Bioinformatics, Peking University	 Data and Information Management Group, Peking University

 Dataverse Project About ▾ Community ▾ Best Practices ▾ Software ▾ Contact

# The Dataverse Project



## Open source research data repository software

Enjoy full control over your data. Receive *web visibility*, *academic credit*, and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

 Researchers

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data*. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

 Journals

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

 Institutions

  
复旦大学社会科学数据平台

首 页 新闻动态 国际资源 代言系列 平台简介 相关链接 联系我们 帮 助

复旦大学社会科学数据平台

复旦大学社会科学数据平台：收集、整理和开发中国社会经济发展数据，为学者提供具有最具竞争力的研究条件和数据服务，为学生提供更加坚实的社会科学课堂方法和应用的训练，鼓励跨学科的研究，为复旦大学履行大学传承文明、记录文明的职责和成为国家智库、提供重要的和基础性的支撑。

显示已发布的数据集

已发布的数据集

数据类型

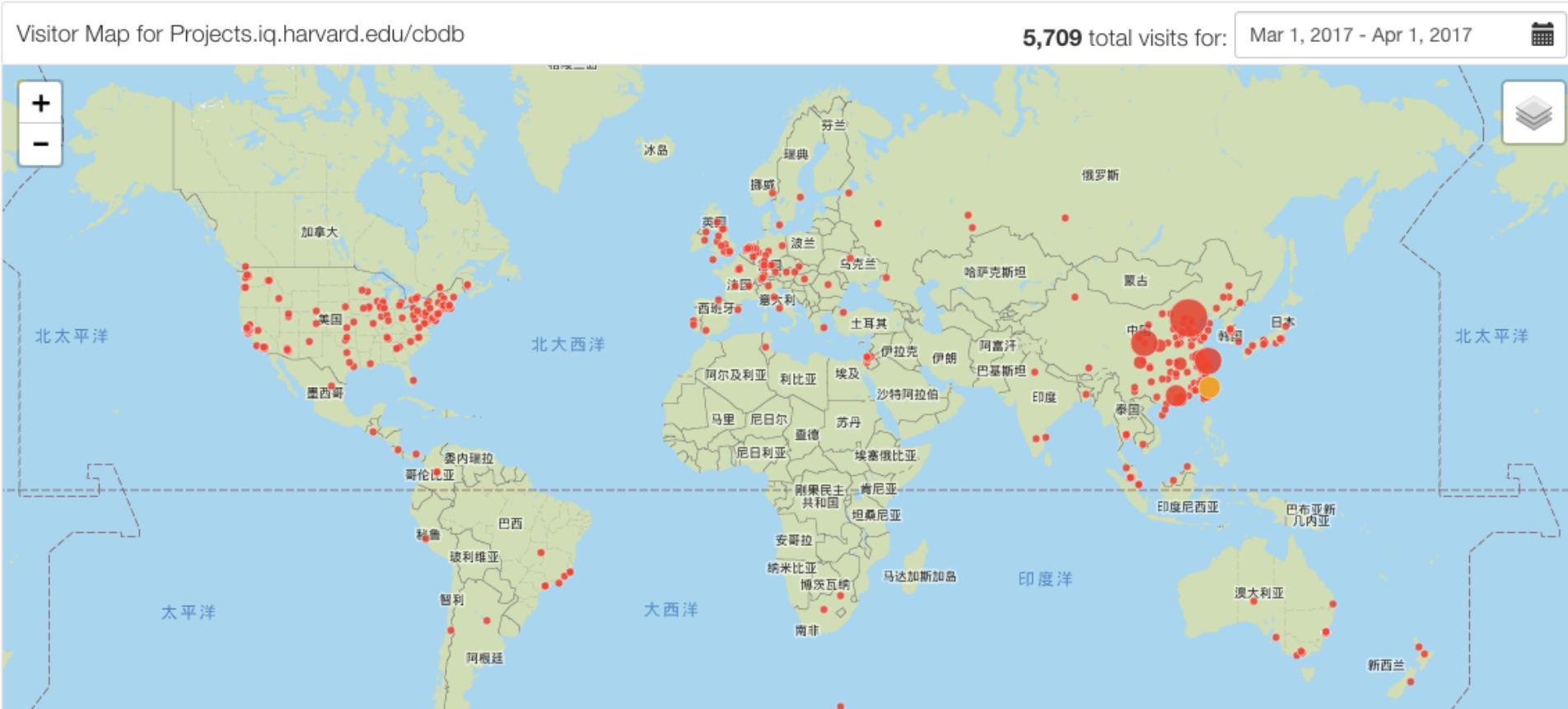
全部 [A] [B] [C] [D] [E] [F] [G] [H] [I] [J] [K] [L] [M] [N] [O] [P] [Q] [R] [S] [T] [U] [V] [W] [X] [Y] [Z]

数据集: 67 | 题目: 683 | 文件: 2,121

名称	机构	发布日期	活跃度
复旦大学能源与环境因子数据库	复旦大学能源研究中心	2014-12-16	■■■■
长三角社会变迁调查	复旦大学社会科学数据中心	2014-12-9	■■■■
数据监护研究与应用	复旦大学文獻信息中心	2015-12-14	■■■■
DIV模型组	复旦大学社会科学数据研究中心	2015-1-12	■■■■
王桂新	复旦大学	2013-3-18	■■■■

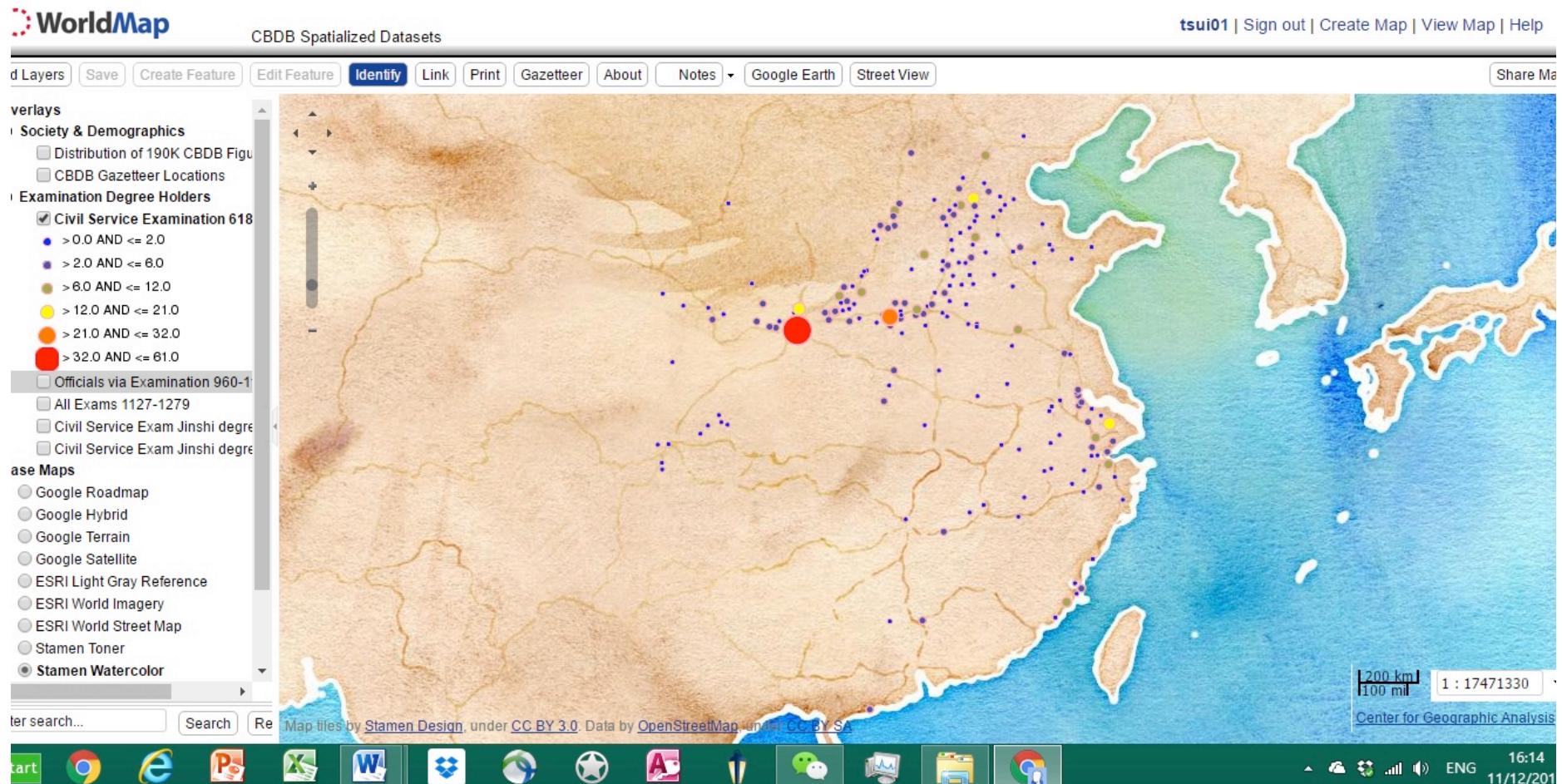
# Data Repositories and their standards

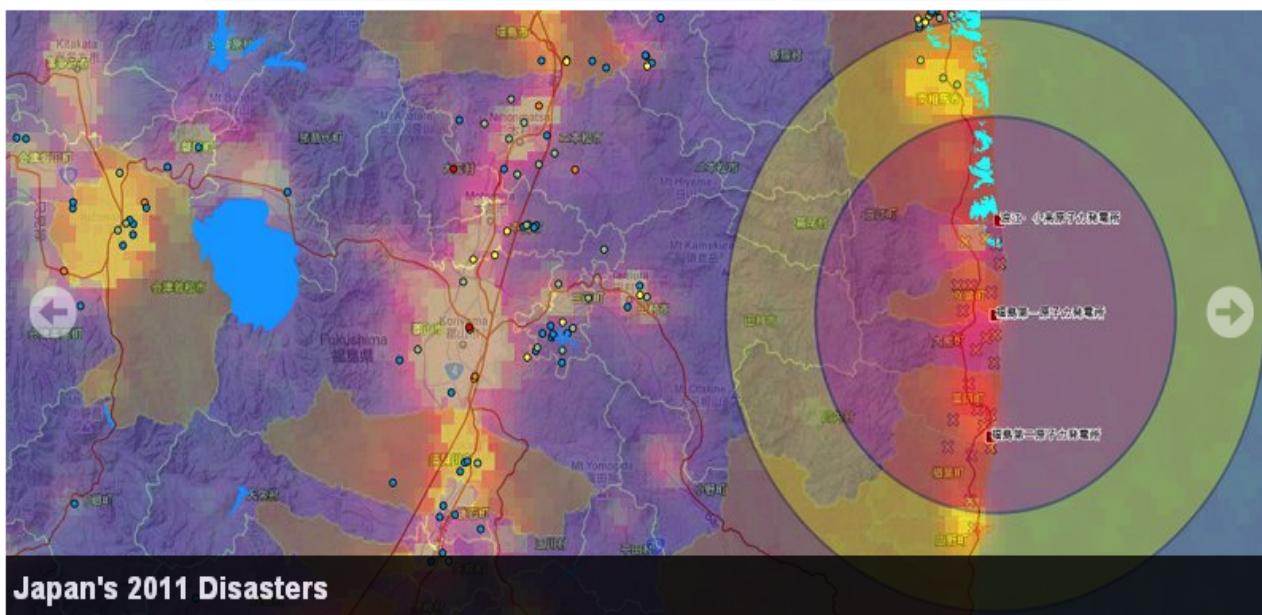
# Tracking user stats for optimizing database user experiences



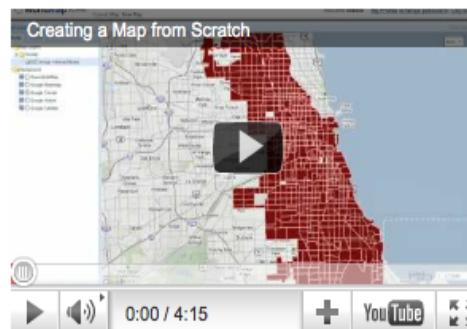
# GIS data platform: WorldMap

## e.g. ChinaMap etc.



[Create a Map](#)[View a Map](#)[About](#)

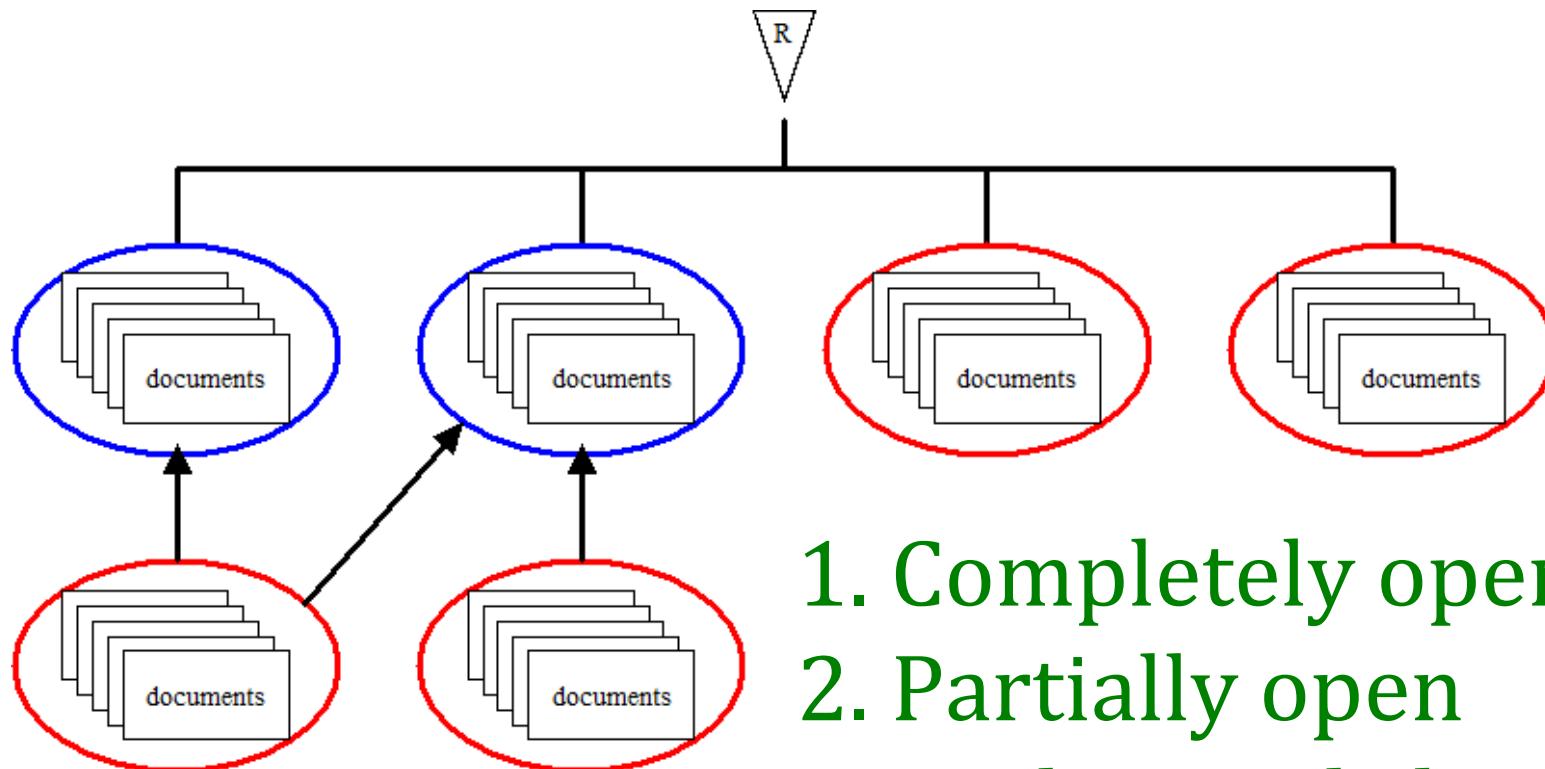
Build your own mapping portal and publish it to the world or to just a few collaborators. WorldMap is open source software.



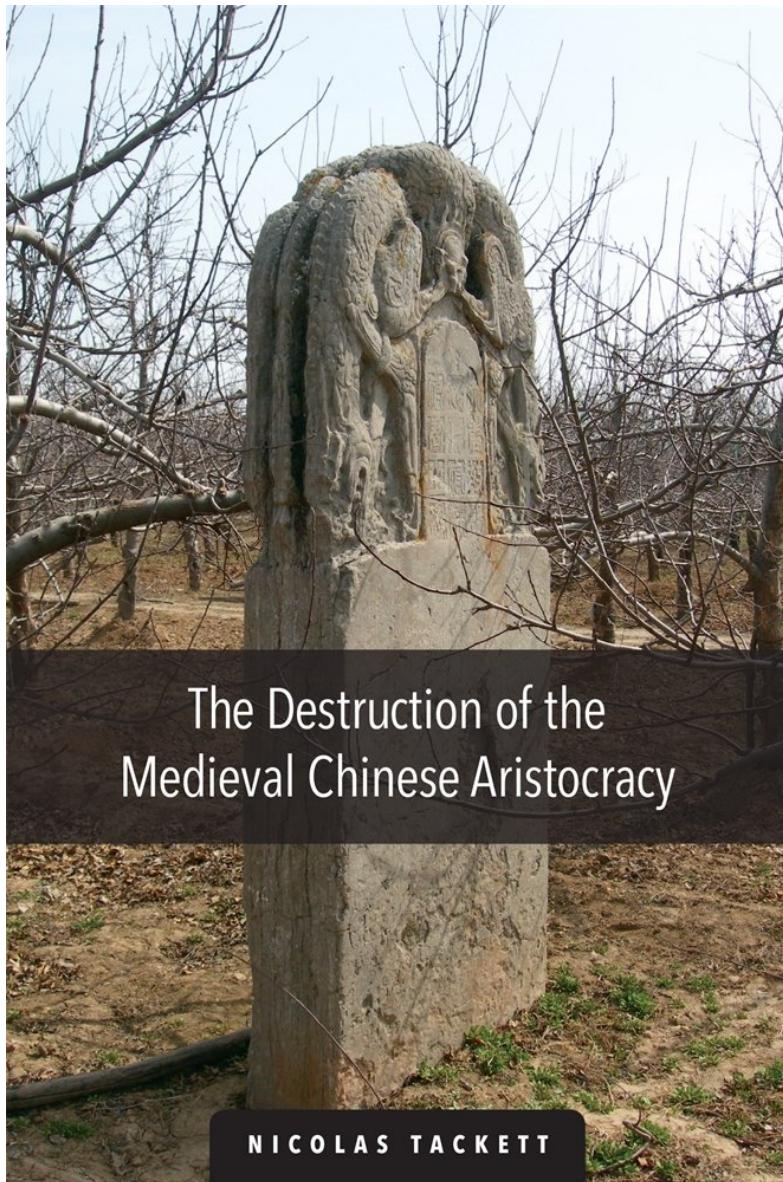
[Watch the WorldMap Quick Start video](#)

<http://worldmap.harvard.edu/>

# Institutional Authorization of Sharing



1. Completely open
2. Partially open
3. Authorized sharing  
esp. full text databases



## The Destruction of the Medieval Chinese Aristocracy

NICOLAS TACKETT

Information, Territory, and Networks

The Crisis and Maintenance of Empire in Song China

Hilde De Weerdt

A photograph of a book cover. The title "Information, Territory, and Networks" is at the top in a large, dark font. Below it, a subtitle "The Crisis and Maintenance of Empire in Song China" is in a smaller, dark font. The author's name, "Hilde De Weerdt", is at the bottom right. The background of the cover is a light beige color. On the right side, there is a vertical column of Chinese characters and a horizontal strip showing a faint illustration of a historical scene.

# Other relevant tools of infrastructure importance

---

1. 跨庫書目檢索系統 Cross-Catalogue Query System of Ancient Chinese Books
2. OCR技術與中文文本資源的開放 OCR Techniques and Textual Resources e.g. CTEXT
3. 標記與可視化工具 Tagging and Visualization Tools e.g. MARKUS
4. 代碼表/權威檔 Code Tables/Authority Files e.g. data in CBDB

# 计算机告诉你，唐朝诗人之间的关系到底什么样？

前进四先生 发表于 2017-03-15 11:28

在我还念中学的时候，每当心情不好，就靠读诗词来排遣，慢慢读得多了，就发现唐朝诗人之间存在着微妙的关系。比如杜甫非常喜欢李白，到了做梦都想见李白的地步：三夜频梦君，情亲见君意（《梦李白》）。而李白向孟浩然表过白：吾爱孟夫子，风流天下闻（《赠孟浩然》）。孟浩然的好基友则是王昌龄：数年同笔砚，兹夕间衾裯（《送王昌龄之岭南》）。

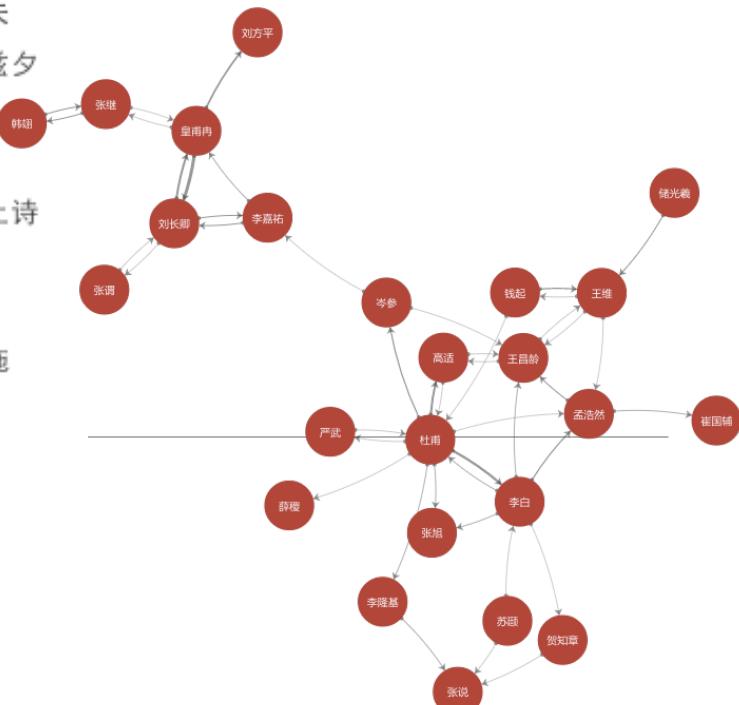
出于好奇心，我一度想理清楚他们之间的关系。但是全唐诗一共四万多首，再加上诗人之间经常称呼对方的别称，整理起来非常麻烦，慢慢的也就绝了这个念头。

前不久又想起来这段十五年前的心事。事不宜迟，拖了这么多年的愿望，不能再拖了。

这次，我将编程完成这件事。

前面已经说过，这件事主要的麻烦在于以下两点：

- 全唐诗数量太多，一共四万多首。
- 诗人的别称太多，比如杜甫：按字称为子美，按排行称为杜二，按官职称为杜工部。



# Communication between participating actors

---

# Web-based Communication of DS/DH actors

---

- **Cyberinfrastructure website**
  - not centrally managed, but ideally generated and updated automatically by crawling/crowdsourcing project updates
  - a marketplace of ideas and resources for collaboration
- **Mailing lists**
  - still a good way to discuss a topic and share one's views in depth
  - e.g. Humanist Discussion Group
- **Social network accounts e.g. Wechat, FB**
  - somewhat restrictive, but users are often very active and willing to help

# <http://dhhumanist.org/>

---

[Home](#) [About](#) [Subscribe](#) [Search](#) [Member Area](#)



«[T]ruth is not born nor is it to be found inside the head of an individual person, it is born *between people* collectively searching for the truth, in the process of their dialogic interaction.... »  
Mikhail Bakhtin, *Problems of Dostoevsky's Poetics*, trans. Caryl Emerson (University of Minnesota Press, 1984, pp. 110).

"We" philosophers are... distinguished ... by our ability to engage in continuous conversation, testing one another, discovering our hidden presuppositions, changing our minds because we have listened to the voices of our fellows. Lunatics also change their minds, but their minds change with the tides of the moon and not because they have listened, really listened, to their friends' questions and objections. Montaigne in his tower and Kierkegaard in his isolation are of that goodly listening company, despite their solitude. The inner voices that they heard were real enough, Montaigne remembering his friend Etienne de la Boetie, and Kierkegaard mocking Pastor Adler. Amelie Oksenberg Rorty, *Experiments in Philosophic Genre: Descartes' "Meditations"*, *Critical Inquiry* 9.3 (1983): 562.

---

*Humanist* is an international seminar on digital humanities founded in 1987. Its aim is to provide a forum for discussion of intellectual, scholarly, pedagogical, and social issues and for exchange of information among participants. *Humanist* is a publication of the Alliance of Digital Humanities Organizations ([ADHO](#)) and an affiliated publication of the American Council of Learned Societies ([ACLS](#)). For more information on the activities of the world-wide digital humanities community and for ways to get involved see the [ADHO website](#). To apply for membership, click on 'Subscribe', above.



Wednesday 1:43 PM



高瑾-伦敦大学学院



高瑾-伦敦大学学院

这些书也是研究生入学必读书单

Thursday 1:37 PM



徐力恒-哈佛大學CBDB

@高瑾-伦敦大学学院 有沒有興趣分享一下研究生讀DH的體會？



高瑾-伦敦大学学院

好呀～群里还有很多DH毕业研究生呢 😊



徐力恒-哈佛大學CBDB

邀請幾位同學，在零壹Lab上每人寫一篇短文，會對發展DH教學很有參考價值呢！



## 中國歷代人物傳記資料庫（CBDB）線上課程

科技部數位人文籌畫小組 • 13 videos • 358 views • Last updated on Dec 4, 2015

對文史研究者而言，中國歷代人物傳記資料庫（China Biographical Database）是一個十分特殊的數位工具。它並非我們熟悉的史料全文資料庫，在 CBDB 中，我們並不能看到一篇篇的人物傳記。然而，CBDB 擁有許多獨到的查詢功能，彈指之間，即可為諸多問題提供豐富資訊，例如：於南宋寧宗、理宗朝，福建路下的莆田縣共出了多少進士？這些進士彼此間是否有親緣關係？他們在官場上有何互動？他們有哪些共同的友人？透過這些資訊，CBDB 能夠協助研究者重構古人的社交圈、探索古人錯綜複雜的社交網絡。[more](#)

[▶ Play all](#)[◀ Share](#)[+ Save](#)

- 1 [CBDB線上課程：1-1下載與安裝](#)  
by 科技部數位人文籌畫小組

- 2 [CBDB線上課程：2-1「按人查詢」](#)  
by 科技部數位人文籌畫小組

- 3 [CBDB線上課程：2-2「按入仕途徑查詢」](#)  
by 科技部數位人文籌畫小組

- 4 [CBDB線上課程：2-3「官職查詢」](#)  
by 科技部數位人文籌畫小組

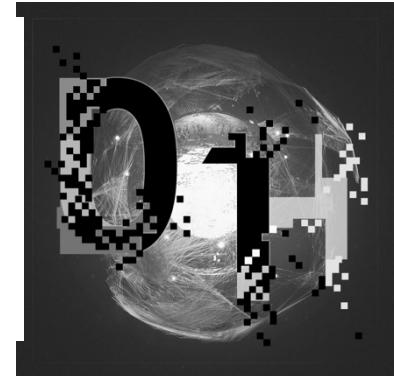
# Active DH scholarly communities esp. for Chinese history



零壹Lab

记录数字媒介之日常

反思科技与人文精神



零壹Lab  
(ID: lingyilab)

# 【交换机】国内首个数字史学课程——独家推送课程资料！

(原创) 2016-12-12 王涛 零壹Lab

## 课程简介

数字工具与世界史研究

主讲：王涛  
南京大学历史学院  
  
上课地点：仙II-301  
时间：2016/17学年  
周二14:00-16:00



“ ”

## 6. 史学成果的网络发布：以Omeka为例

数字史学的研究成果发布超越了印刷载体的束缚，网络写作将成为一个重要的补充。为此，我们将引入元数据（Metadata）的概念，介绍以Dublin Core为代表的元数据标准。

Toolkit技能：Omeka的使用。我们将介绍网络发布平台Omeka的工作原理和相关专业术语，并通过实际的操作了解网络写作与传统写作的差异。

## 7. 史学研究工具I：社会网络分析

我们将讨论社会网络分析（Social Network Analysis，SNA）的what，why以及how。在分析了SNA的学术发展脉络后，我们指出历史研究引入社会网络分析的方法，能够打开新的研究领域。我们将介绍SNA的分析指标，比如路径（Path）、紧密中心度（closeness centrality）、间距中心度（betweenness centrality）等，用它们来分析网络中权力的支配、权力的交换、以及社会威望等问题。

Toolkit技能：Gephi的使用。

## 8. 史学研究工具II：自然语言处理

历史研究的重要史料是文本，在数字史学领域，这里涉及的就是自然语言处理的问题。我们将介绍自然语言的概念，并介绍基于不同算法的自然语言处理方

# 零壹Lab內容：

## <http://www.iwgc.cn/list/11647>

- 统计学与人文研究的结合如何可能？——清华大学统计学研究中心  
邓柯博士访谈
- DH年度大奖接受投票啦！请支持古文标记平台MARKUS码库思！
- 香港中文大学图书馆数码学术研究研讨会
- 你是数字人文者（DHer）么？
- 海量？智慧？整洁？混乱？——人文学科中的数据
- 《图书馆论坛》“数字人文” 专栏征稿启事
- 谁是数字人文的受益者？——台湾DADH数字人文会议最新报道
- 国内首个数字史学课程——独家推送课程资料！
- 美国数字人文学者推介：霍伊特·朗(Hoyt Long，芝加哥大学)
- 独家专访Susan Schreibman：如果你习惯待在某个固定领域做研究，  
那最好别选数字人文

Facebook : <https://www.facebook.com/DHVirtualLab>

# 數位人文實驗室

DIGITAL HUMANITIES VIRTUAL LAB



徐力恒

首頁 3



訊息



消息

?

粉絲專頁

訊息

通知

已說讚 ▾ 追蹤中 ▾ 分享 ...

發送訊息

# 數位人文實驗室 Digital Humanities Virtual Lab

@DHVirtualLab

首頁

關於

相片

按讚分析

活動

影片

網誌

貼文

服務內容

優惠



數位人文實驗室 Digital Humanities Virtual Lab

由 Ping-tzu Chu 發佈 [?] · 3月13日 23:59 ·

歡迎在花蓮地區的朋友來參加，其它地區當然也歡迎，只是無法提供旅費。

<http://dhit.kktix.cc/events/corpro2017ndhu...>

## 【庫博】Corpro文本分析工作坊

庫博是由闞河嘉老師所開發的文本分析軟體，適合分析大量文本，在很短的時間內就可以對語料庫做出一些有數量分析的觀察。歡迎大...

DHT.KKTIX.CC

已觸及172名用戶

加強推廣貼文



讚



留言



分享



Yu Jung Cheng 、 Jing-xin Yen 和其他 2 人



留言.....



顯示全部



Digital WISTM

教育



Digital chumchon lantanoi

教育



Digital Human Library



Fordham Graduate Student Di...



Digital Media &amp; Learning Rese...

中文(台灣) · 中文(简体) · English (US) ·  
Español · Português (Brasil)

+



TADH

# 臺灣數位人文學會

Taiwanese Association for Digital Humanities





## 课程目标

- 掌握数字人文研究的基本概念和研究状况，并知道关注领域新进展的渠道；
- 了解重要的数字化资源和工具（尤其是针对中国文史研究的工具，如CBDB），知道利用的方法，并提高动手和解决操作问题的能力；
- 获得通过数据思考学术问题的能力，能把问题部分地转化为数字化手段能分析和呈现的课题，并摸索如何建立对自身有用的数据集；
- 了解数字人文研究的成果和新范式下学术成果的形态，并具备批判眼光，反思其研究方法和结论。

1. 数字人文的现状、基本概念和理论
2. 关系型史学数据库（上）：从用户角度看CBDB
3. 关系型史学数据库（下）：从开发者角度看CBDB
4. 电子地图和地理空间分析
5. “中国R会议”（清华大学）
6. 社会网络分析
7. “第二届北京大学数字人文论坛”（北大图书馆）
8. 文本的处理、提取和标记：MARKUS和VISUS
9. 数字人文范式下的版本目录学和书籍史
10. 总结和成果研讨



## // Blog Post

**UoN Blogs / China Policy Institute Blog**

June 13, 2016, by [Editor](#)

# The Digital Humanities as an Emerging Field in China

Written by [Lik Hang Tsui](#).



The "digital humanities" (usually translated as *shuzi renwen* 数字人文 in mainland China and *shuwei renwen* 數位人文 in Taiwan) have recently received a lot of attention in Chinese academic circles, even though it took a long time for the concept to come to the attention of mainland China universities. The first digital humanities centre in China was established by [Wuhan University](#) in 2011. It remains the only mainland Chinese member of [centerNet](#), an international network of digital humanities research centres.

But even though such research centres are so rare in China, plenty of Chinese scholars have been taking part in the field and have been conducting digital research for more than a decade. Several key universities have academic departments and centres that are developing digital projects for the humanities, such as

A cyberinfrastructure would provide the mechanisms and systems (e.g. APIs) for us to collaborate in digital scholarship for Chinese history.

It is now time for this conversation to begin for our field.

---

Thank you!

[tsui01@fas.harvard.edu](mailto:tsui01@fas.harvard.edu)