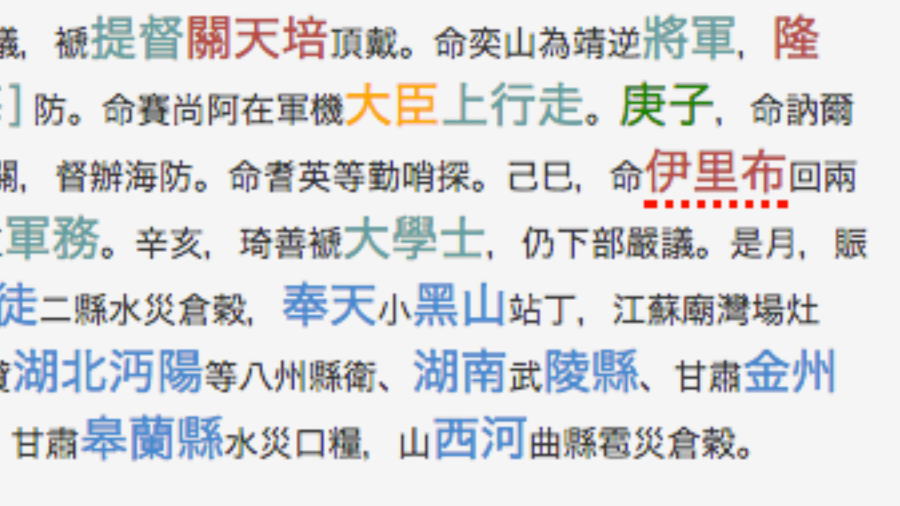
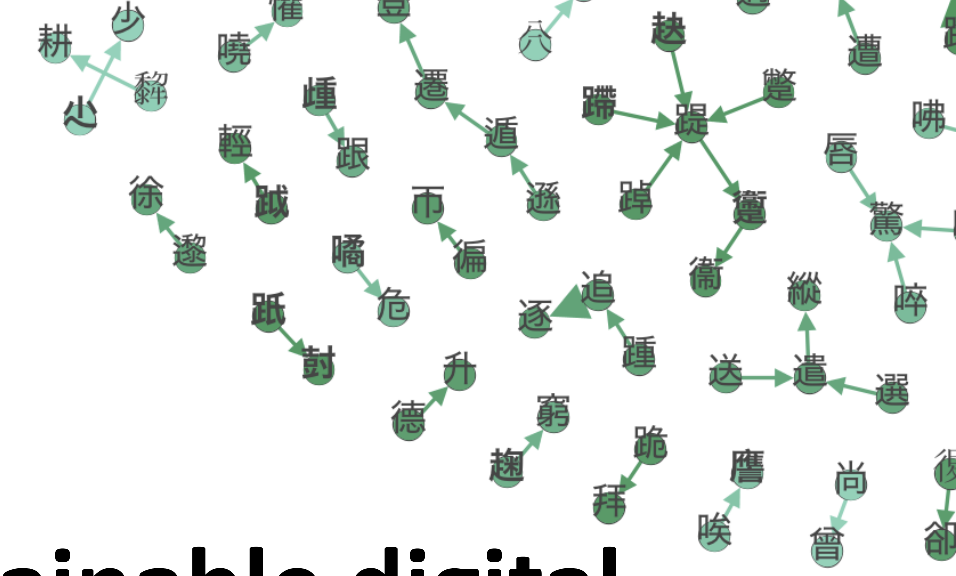


義，䙬提督關天培頂戴。命奕山為靖逆將軍，隆
] 防。命賽尚阿在軍機大臣上行走。庚子，命訥爾
關，督辦海防。命耆英等勤哨探。己巳，命伊里布回兩
軍務。辛亥，琦善䙬大學士，仍下部嚴議。是月，賑
徒二縣水災倉穀，奉天小黑山站丁，江蘇廟灣場灶
湖北沔陽等八州縣衛、湖南武陵縣、甘肅金州
甘肅皋蘭縣水災口糧，山西河曲縣雹災倉穀。



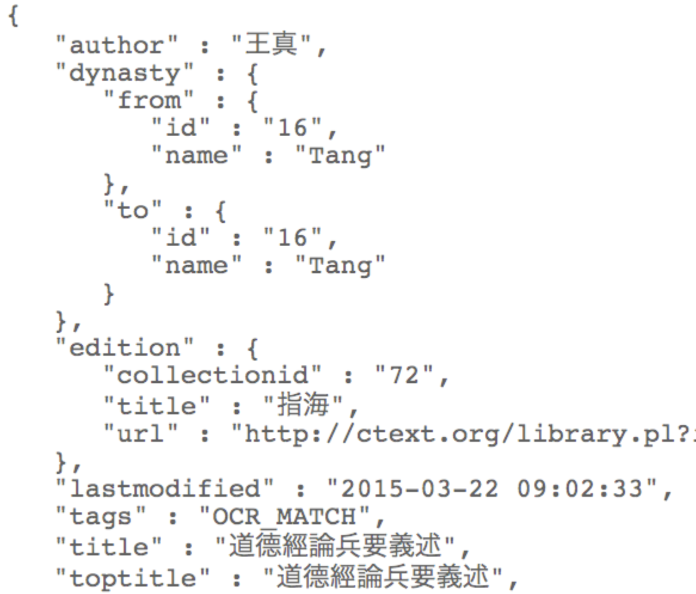
Towards a sustainable digital infrastructure for historical Chinese texts

Donald Sturgeon

Fairbank Center for Chinese Studies

Harvard University

sturgeon@fas.harvard.edu



```
{
   "author" : "王真",
   "dynasty" : {
      "from" : {
         "id" : "16",
         "name" : "Tang"
      },
      "to" : {
         "id" : "16",
         "name" : "Tang"
      }
   },
   "edition" : {
      "collectionid" : "72",
      "title" : "指海",
      "url" : "http://ctext.org/library.pl?
   },
   "lastmodified" : "2015-03-22 09:02:33",
   "tags" : "OCR_MATCH",
   "title" : "道德經論兵要義述",
   "toptitle" : "道德經論兵要義述",
```

Chinese Text Project (ctext.org)

- Largest full-text digital library of pre-modern Chinese
  - 25 million pages of scanned primary source texts
  - 5 billion characters of transcription
  - Used daily by 25,000 people

- 2005-2013: primarily *static* database
- 2013-present: major *dynamic* components

- Originally: centrally edited and maintained
- Now: primarily maintained by crowdsourcing

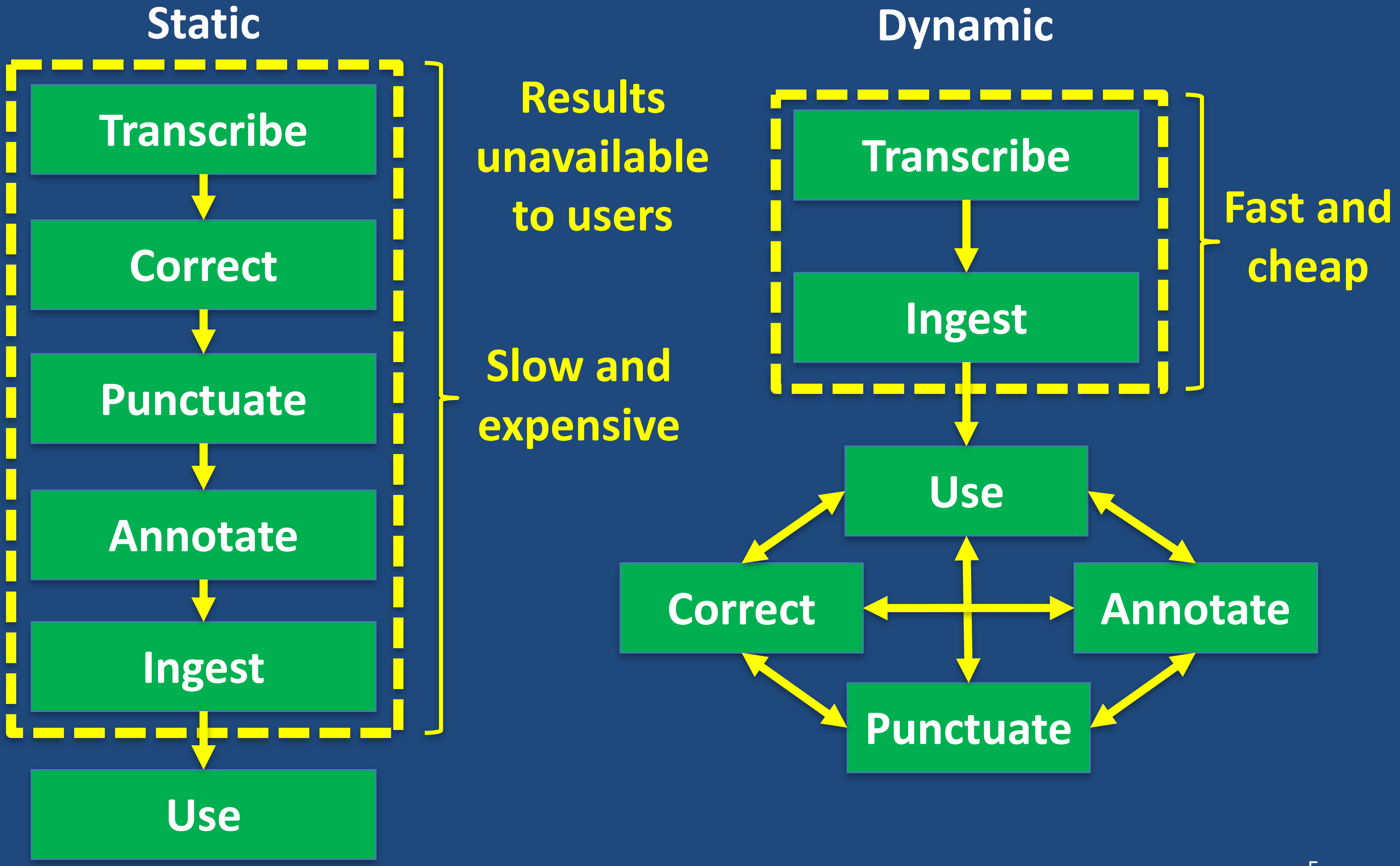
ctext.org as digital infrastructure

- Key components:
  - Optical Character Recognition
  - Crowdsourcing system
  - Application Programming Interface
- Example use cases:
  - Searchable database of texts
  - Scalable transcription tool
  - Repository for data mining

Digitization

- Scanning historical materials *relatively* cheap
  - Established technology
  - Storage & processing costs decrease exponentially
- Scanned images for conservation of material
  - Access to material (e.g. online image browsers)
  - No risk of damage to rare or unique objects
- Large-scale scanning projects undertaken
  - E.g. Harvard-Yenching: 5 million+ pages
  - Even larger scanning projects in Mainland China

Static versus dynamic



Static

Transcribe

Correct

Punctuate

Annotate

Ingest

Use

Results unavailable to users

Slow and expensive

Dynamic

Transcribe

Ingest

Fast and cheap

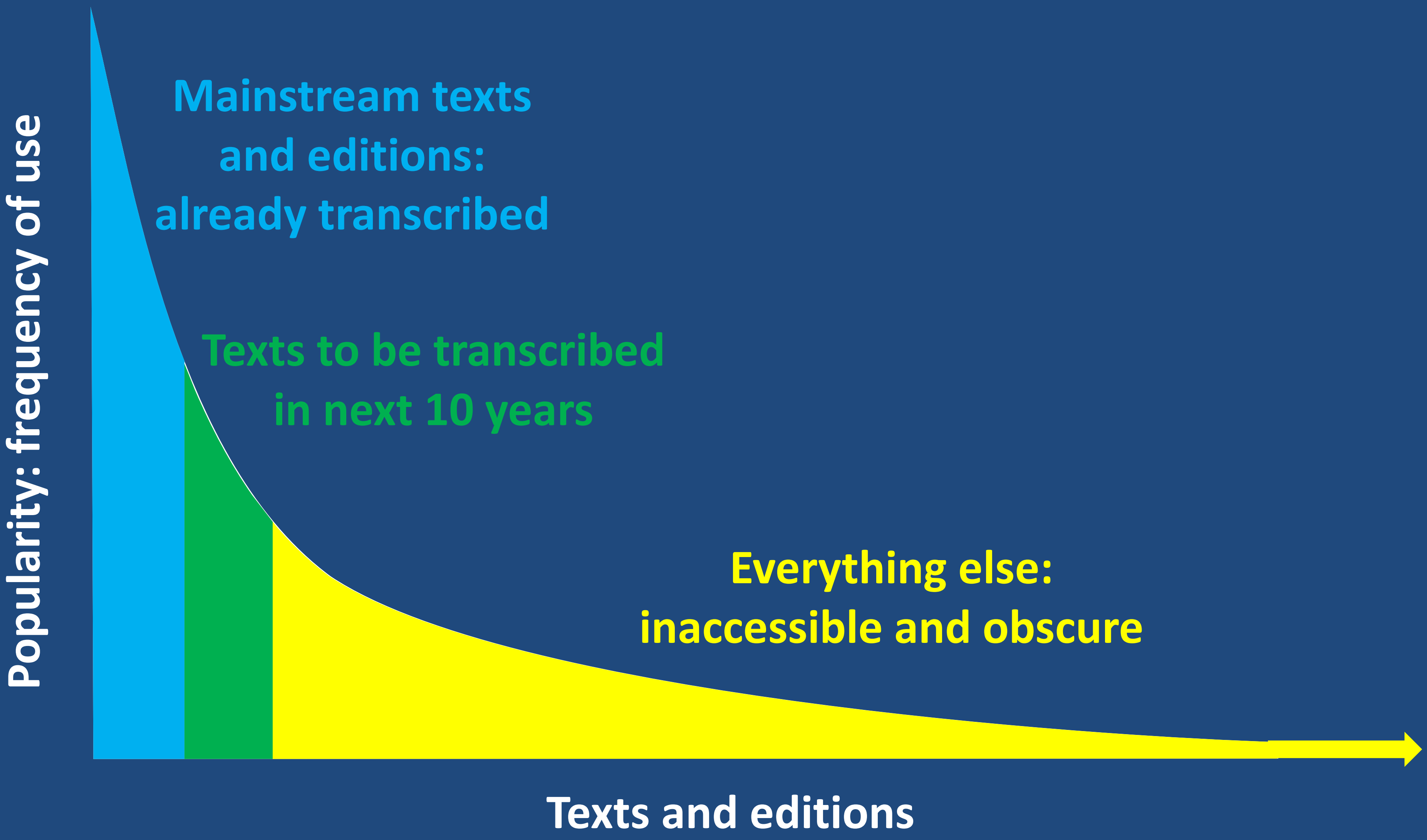
Use

Correct

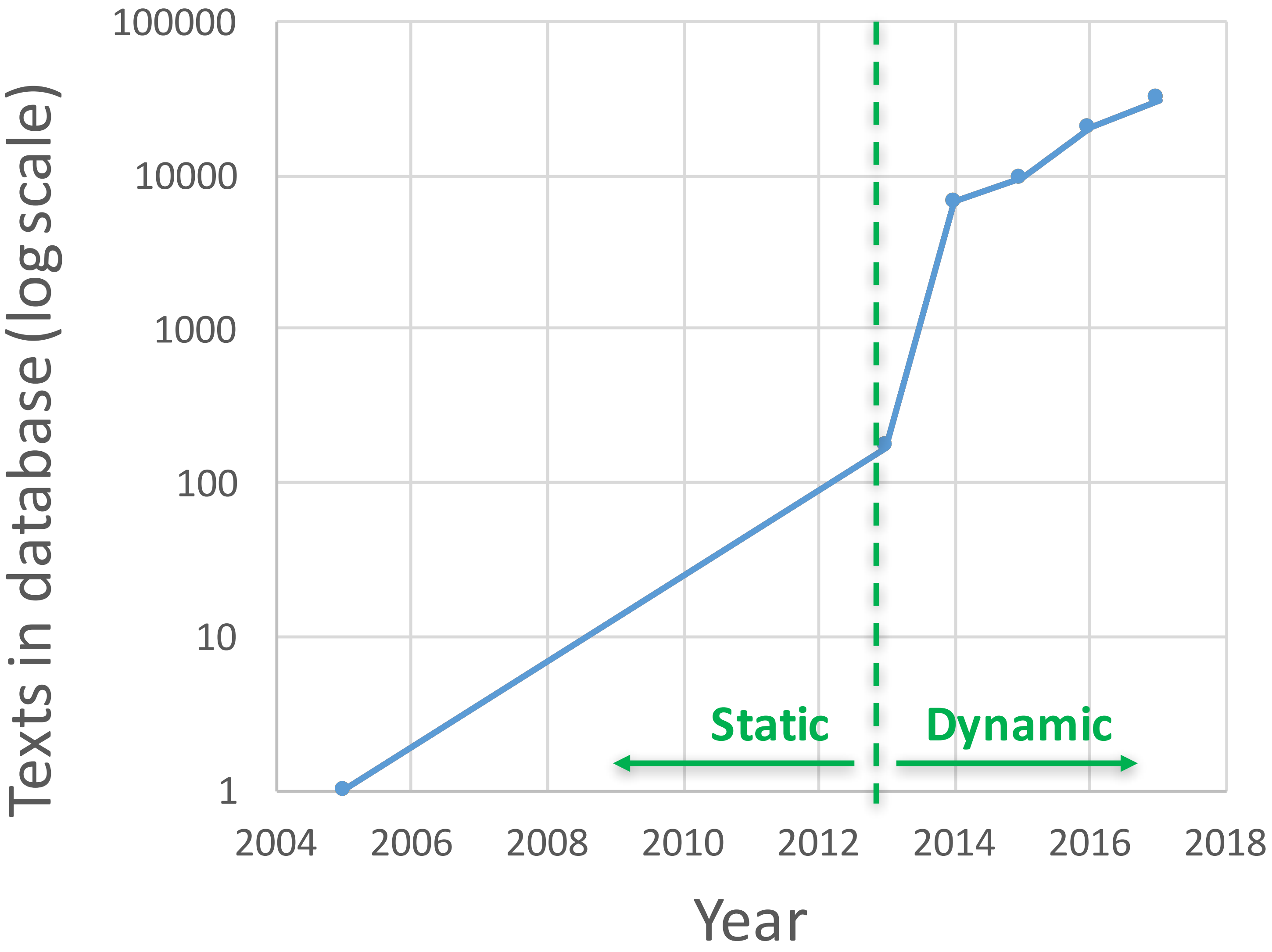
Annotate

Punctuate

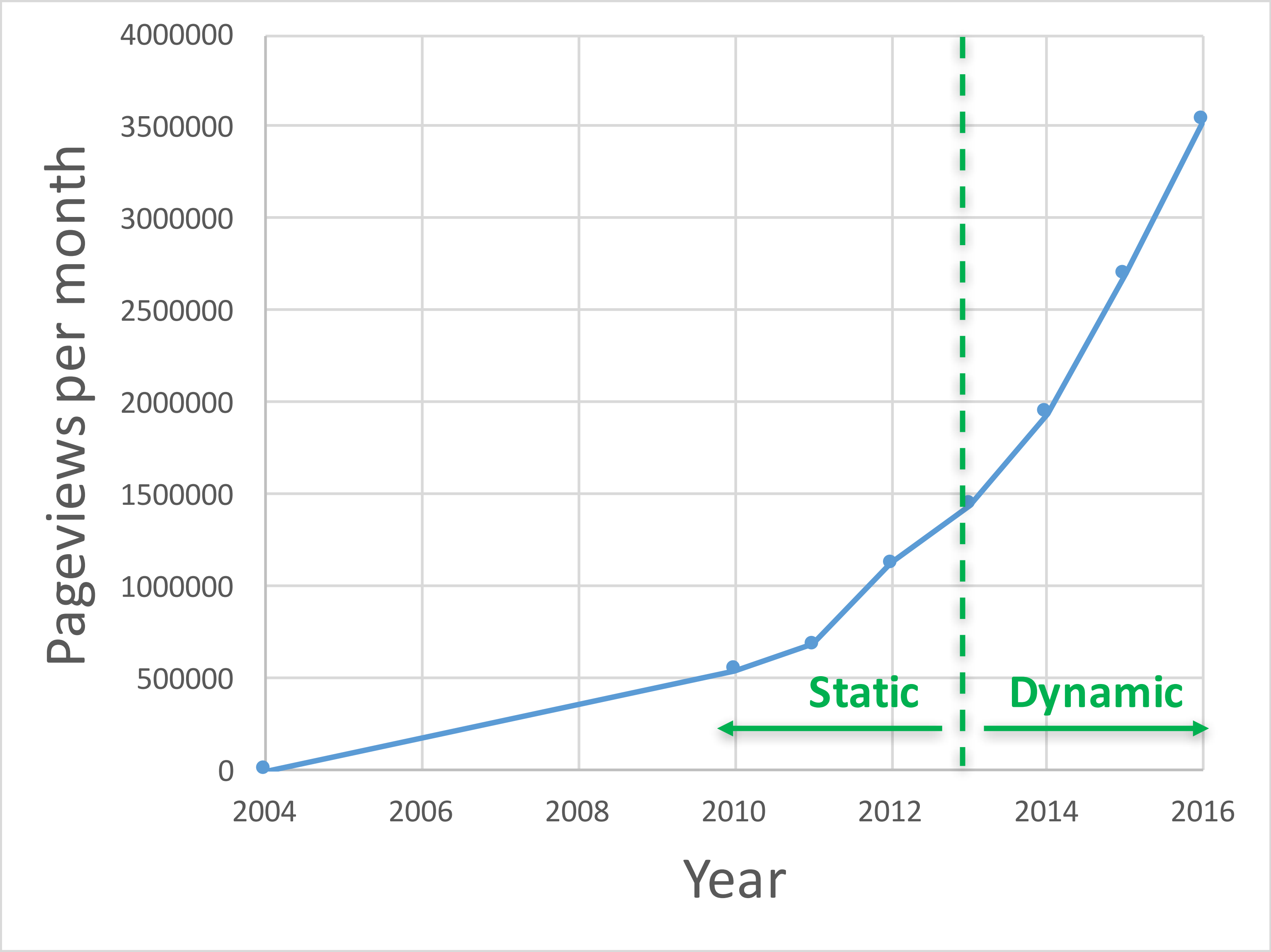
The Long Tail



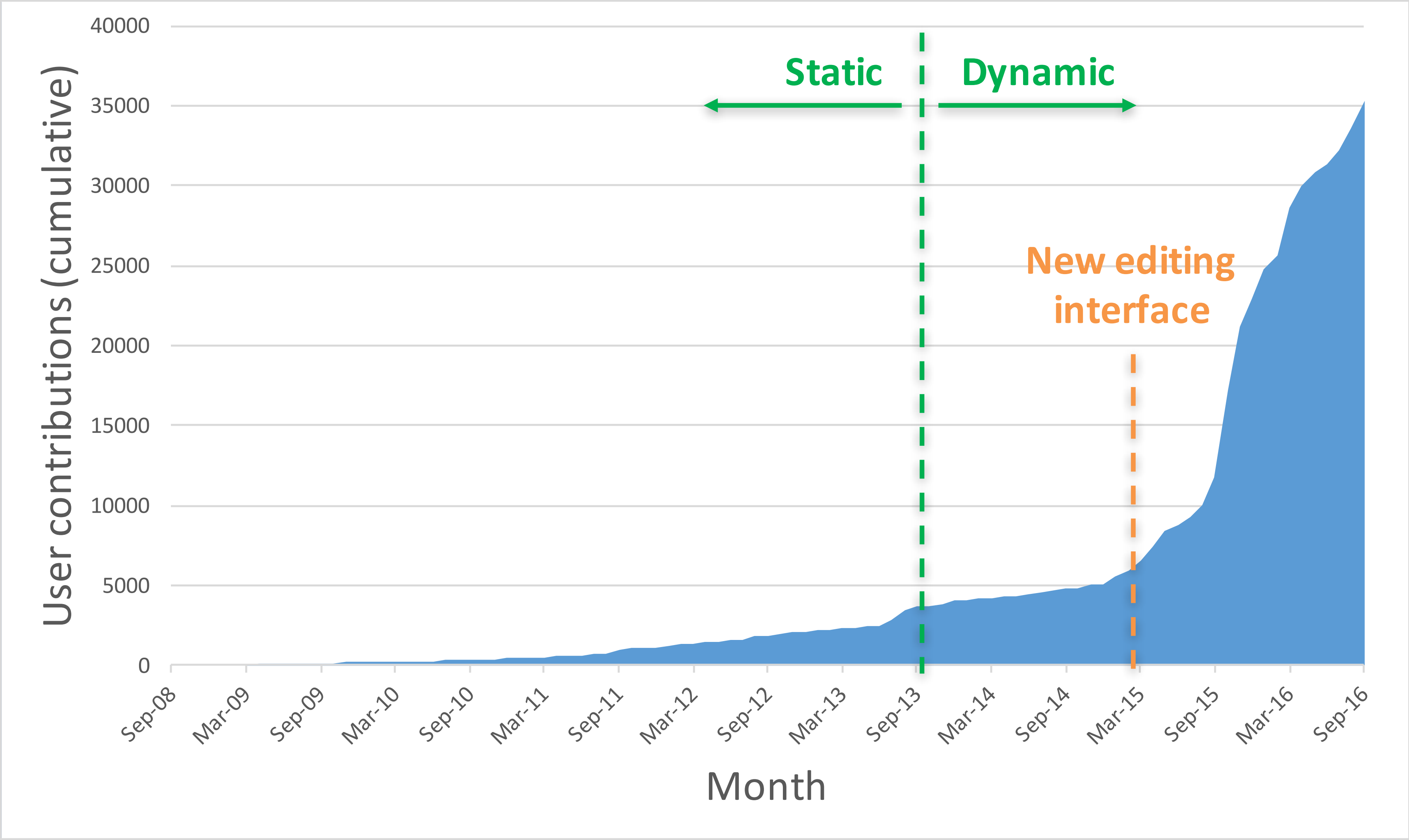
Database size



Database usage



User contributions

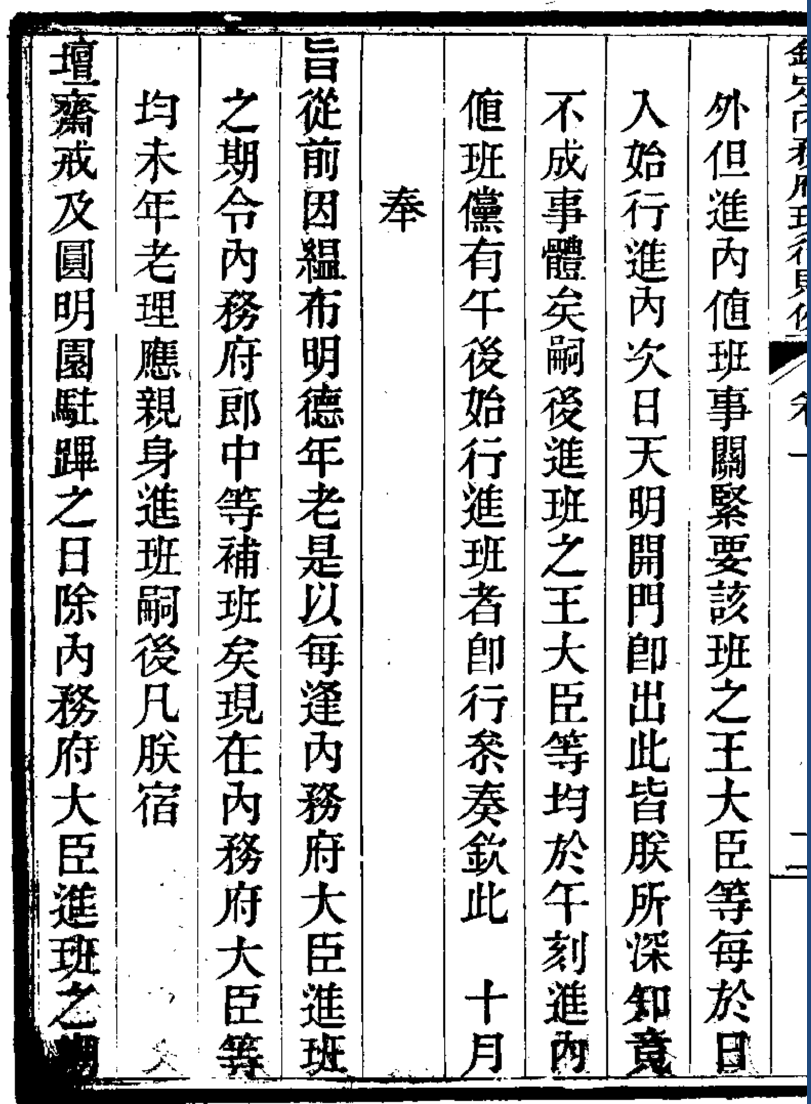


Domain-specific OCR

- Leverage domain knowledge
  - Language, writing conventions
  - Existing transcriptions, text reuse, other editions
- Easier problem => more accurate results
  - Compare with e.g. Google Books OCR
- Enables image search & basis for transcription
  - Frequent use => more opportunities for correction

ctext.org vs Google Books

Google books scan



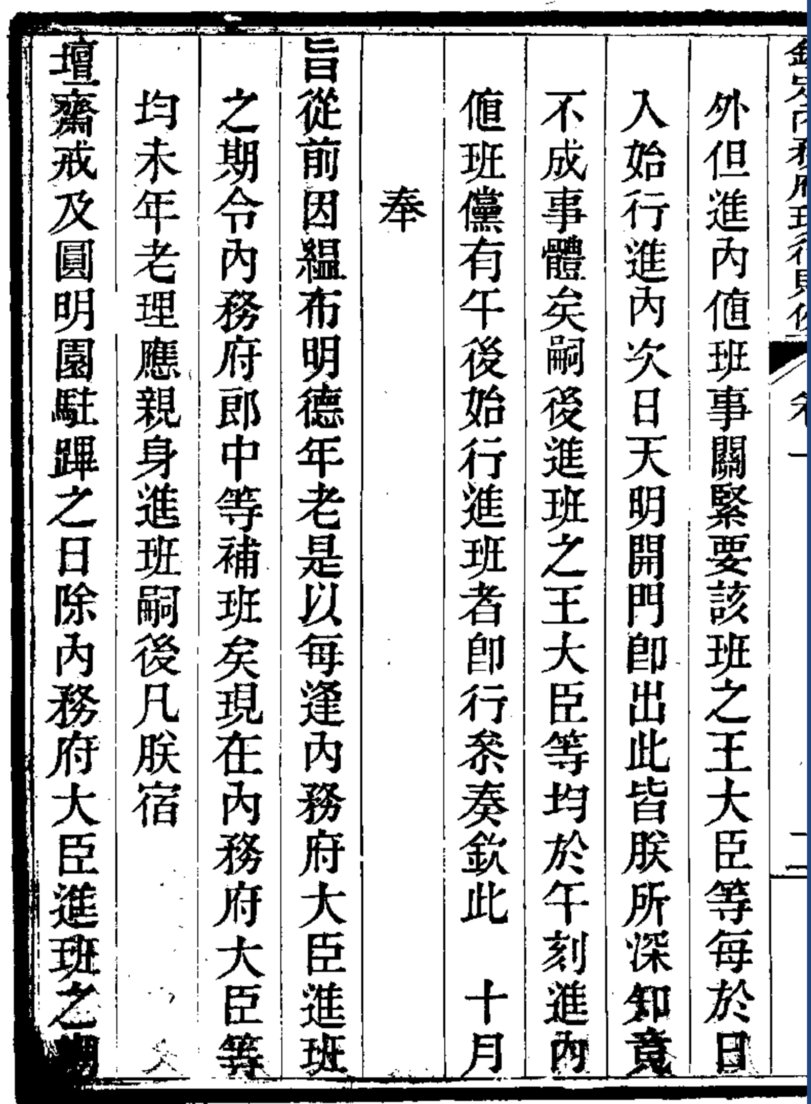
外但進內值班事關緊要該班之王大臣等每於日
入始行進內次日天明開門卽出此皆朕所深知竟
不成事體矣嗣後進班之王大臣等均於午刻進內
值班儻有午後始行進班者卽行叅奏欽此　十月
奉
旨從前因緼布明德年老是以每逢內務府大臣進班
之期令內務府郎中等補班矣現在內務府大臣等
均未年老理應親身進班嗣後凡朕宿
壇齋戒及圓明園駐蹕之日除內務府大臣進班之期

Google books OCR result

一外佴進內他珙事關緊要該班之王大臣等每於日1.
不成事體矣嗣後進班之王大臣等均^午刻進內
進値班儻有午後^行進賽者卽行^溱桀此
旨從前因媪布明德年老是^每逢內務府大臣
4步^13力,X^^,^II顷丙值变暴^^崔內.

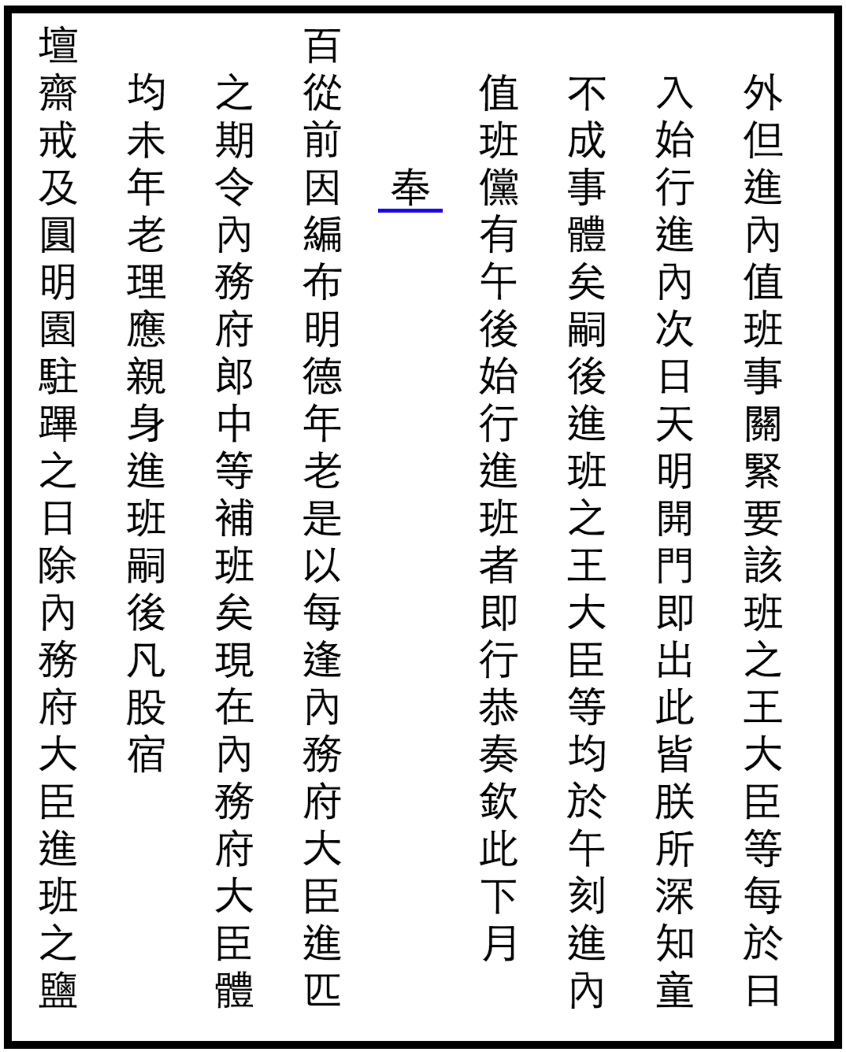
ctext.org vs Google Books

Google books scan



外但進內值班事關緊要該班之王大臣等每於日
入始行進內次日天明開門即出此皆朕所深知竟
不成事體矣嗣後進班之王大臣等均於午刻進內
值班儻有午後始行進班者即行叅奏欽此　十月
奉
旨從前因緼布明德年老是以每逢內務府大臣進班
之期令內務府郎中等補班矣現在內務府大臣等
均未年老理應親身進班嗣後凡朕宿
壇齋戒及圓明園駐蹕之日除內務府大臣進班之期

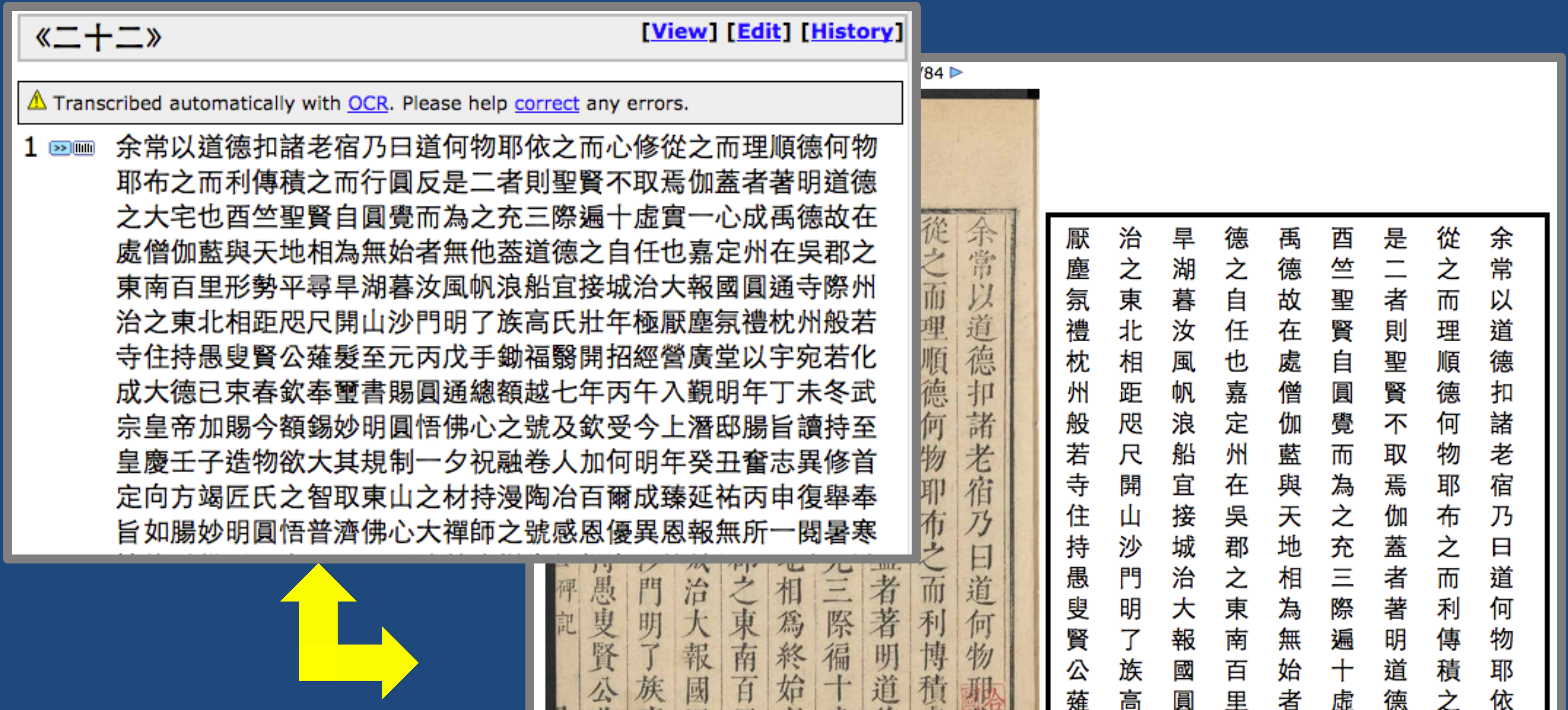
ctext.org OCR result (unedited)



外但進內值班事關緊要該班之王大臣等每於日
入始行進內次日天明開門即出此皆朕所深知童
不成事體矣嗣後進班之王大臣等均於午刻進內
值班儻有午後始行進班者即行恭奏欽此下月
奉
百從前因編布明德年老是以每逢內務府大臣進匹
之期令內務府郎中等補班矣現在內務府大臣體
均未年老理應親身進班嗣後凡朕宿
壇齋戒及圓明園駐蹕之日除內務府大臣進班之鹽

Dynamic approach: advantages

- Imperfect data can still be useful
  - E.g. OCR-derived text enables full-text search



«二十二»

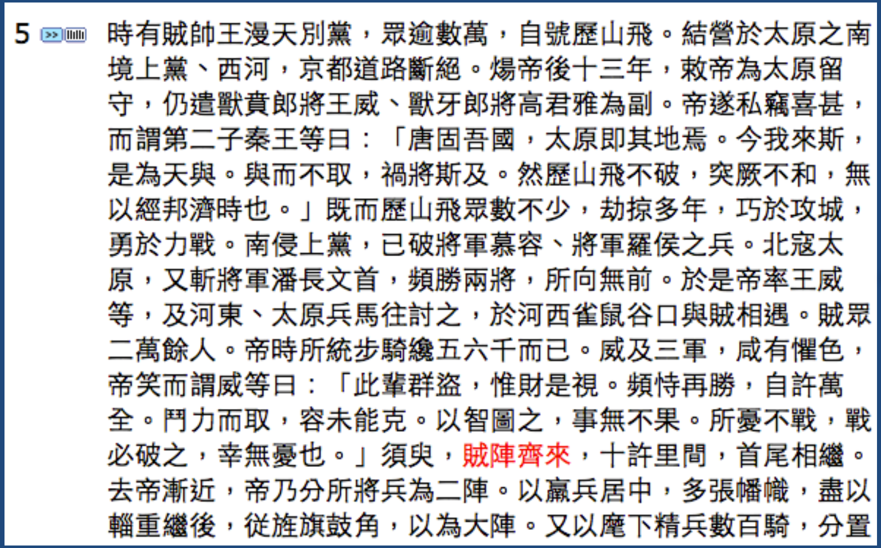
⚠ Transcribed automatically with OCR. Please help correct any errors.

1 余常以道德扣諸老宿乃日道何物耶依之而心修從之而理順德何物耶布之而利傳積之而行圓反是二者則聖賢不取焉伽蓋者著明道德之大宅也西竺聖賢自圓覺而為之充三際遍十虛實一心成禹德故在處僧伽藍與天地相為無始者無他蓋道德之自任也嘉定州在吳郡之東南百里形勢平尋旱湖暮汝風帆浪船宜接城治大報國圓通寺際州治之東北相距咫尺開山沙門明了族高氏壯年極厭塵氛禮枕州般若寺住持愚叟賢公薙髮至元丙戌手鋤福翳開招經營廣堂以宇宛若化成大德已束春欽奉璽書賜圓通總額越七年丙午入覲明年丁未冬武宗皇帝加賜今額錫妙明圓悟佛心之號及欽受今上潛邸賜旨讀持至皇慶壬子造物欲大其規制一夕祝融卷人加何明年癸丑奮志異修首定向方竭匠氏之智取東山之材持漫陶冶百爾成臻延祐丙申復舉奉旨如賜妙明圓悟普濟佛心大禪師之號感恩優異恩報無所一閱暑寒

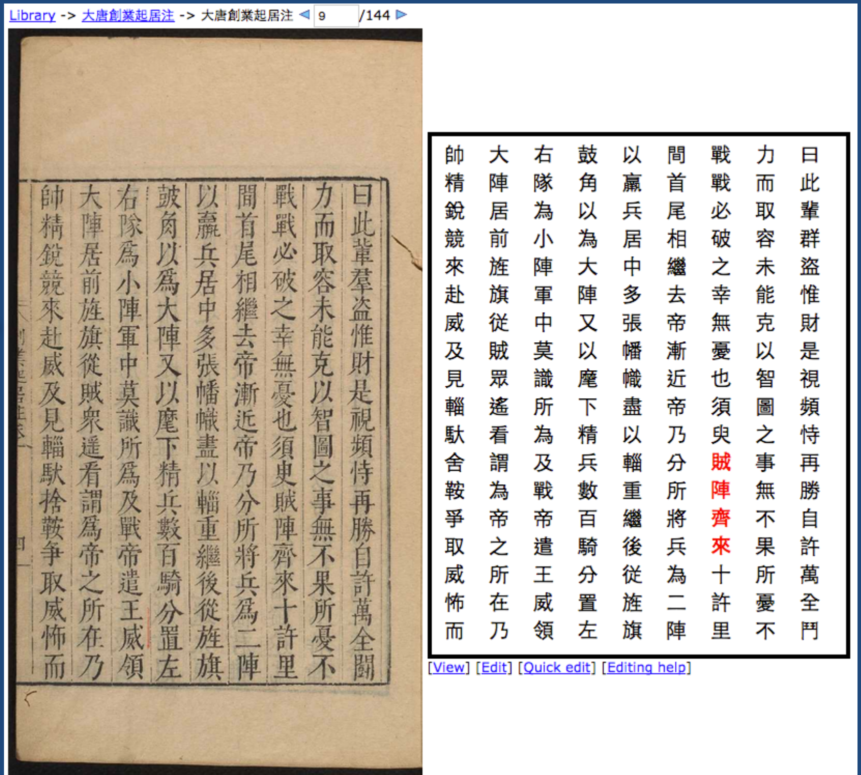
余常以道德扣諸老宿乃日道何物耶依
從之而理順德何物耶布之而利傳積之
是二者則聖賢不取焉伽蓋者著明道德
西竺聖賢自圓覺而為之充三際遍十虛
禹德故在處僧伽藍與天地相為無始者
德之自任也嘉定州在吳郡之東南百里
旱湖暮汝風帆浪船宜接城治大報國圓
治之東北相距咫尺開山沙門明了族高
厭塵氛禮枕州般若寺住持愚叟賢公薙

Dynamic approach: advantages

- Scans can be matched to existing transcription
  - Errors limited to minor alignment errors
  - Punctuation and structure preserved



5 時有賊帥王漫天別黨，眾逾數萬，自號歷山飛。結營於太原之南境上黨、西河，京都道路斷絕。煬帝後十三年，敕帝為太原留守，仍遣獸賁郎將王威、獸牙郎將高君雅為副。帝遂私竊喜甚，而謂第二子秦王等曰：「唐固吾國，太原即其地焉。今我來斯，是為天與。與而不取，禍將斯及。然歷山飛不破，突厥不和，無以經邦濟時也。」既而歷山飛眾數不少，劫掠多年，巧於攻城，勇於力戰。南侵上黨，已破將軍慕容、將軍羅侯之兵。北寇太原，又斬將軍潘長文首，頻勝兩將，所向無前。於是帝率王威等，及河東、太原兵馬往討之，於河西雀鼠谷口與賊相遇。賊眾二萬餘人。帝時所統步騎纔五六千而已。威及三軍，咸有懼色，帝笑而謂威等曰：「此輩群盜，惟財是視。頻恃再勝，自許萬全。鬥力而取，容未能克。以智圖之，事無不果。所憂不戰，戰必破之，幸無憂也。」須臾，賊陣齊來，十許里間，首尾相繼。去帝漸近，帝乃分所將兵為二陣。以羸兵居中，多張幡幟，盡以輜重繼後，從旌旗鼓角，以為大陣。又以麾下精兵數百騎，分置



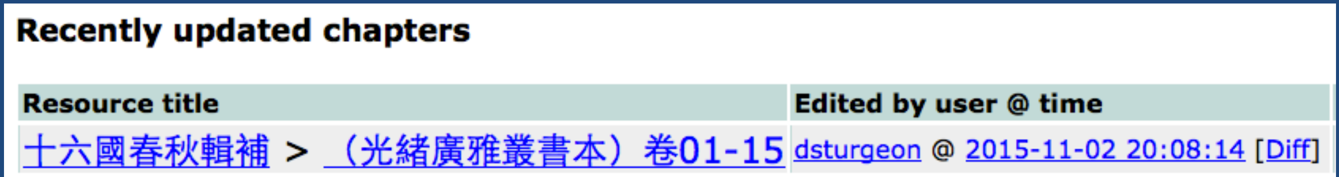
Library -> 大唐創業起居注 -> 大唐創業起居注 ◀ 9 /144 ▶

日此輩群盜惟財是視頻恃再勝自許萬全鬥
力而取容未能克以智圖之事無不果所憂不
戰戰必破之幸無憂也須臾賊陣齊來十許里
間首尾相繼去帝漸近帝乃分所將兵為二陣
以羸兵居中多張幡幟盡以輜重繼後從旌旗
鼓角以為大陣又以麾下精兵數百騎分置左
右隊為小陣軍中莫識所為及戰帝遣王威領
大陣居前旌旗從賊眾遙看謂為帝之所在乃
帥精銳競來赴威及見輜馱舍鞍爭取威怖而

[View] [Edit] [Quick edit] [Editing help]

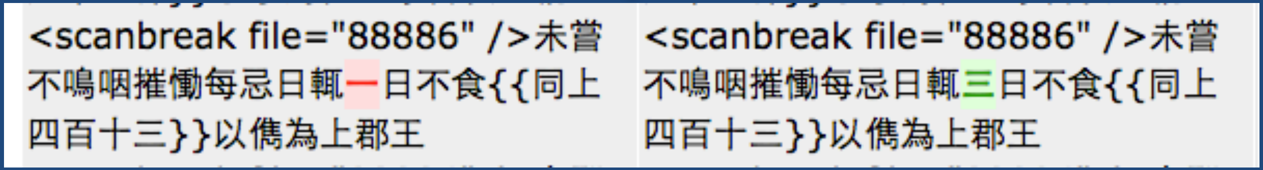
Dynamic approach: challenges

- User-submitted content may be incorrect
  - Need simple way of visualizing & verifying any edit

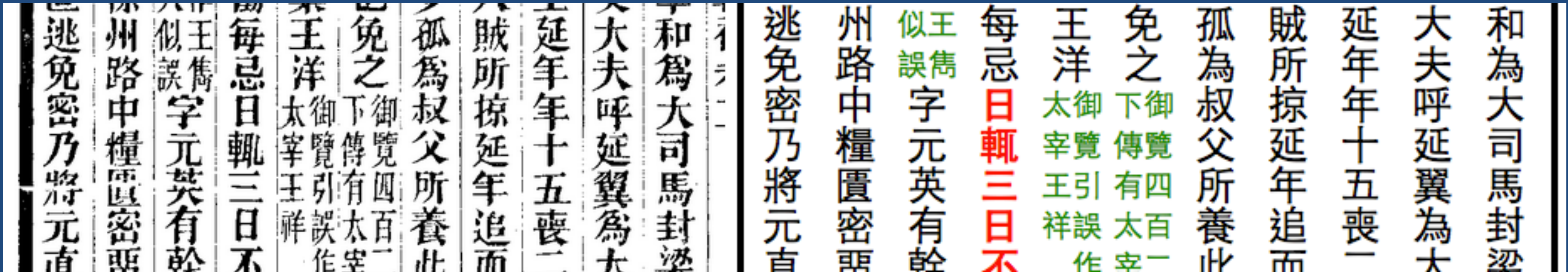


Recently updated chapters

Resource title	Edited by user @ time
十六國春秋輯補 > （光緒廣雅叢書本）卷01-15	dsturgeon @ 2015-11-02 20:08:14 [Diff]



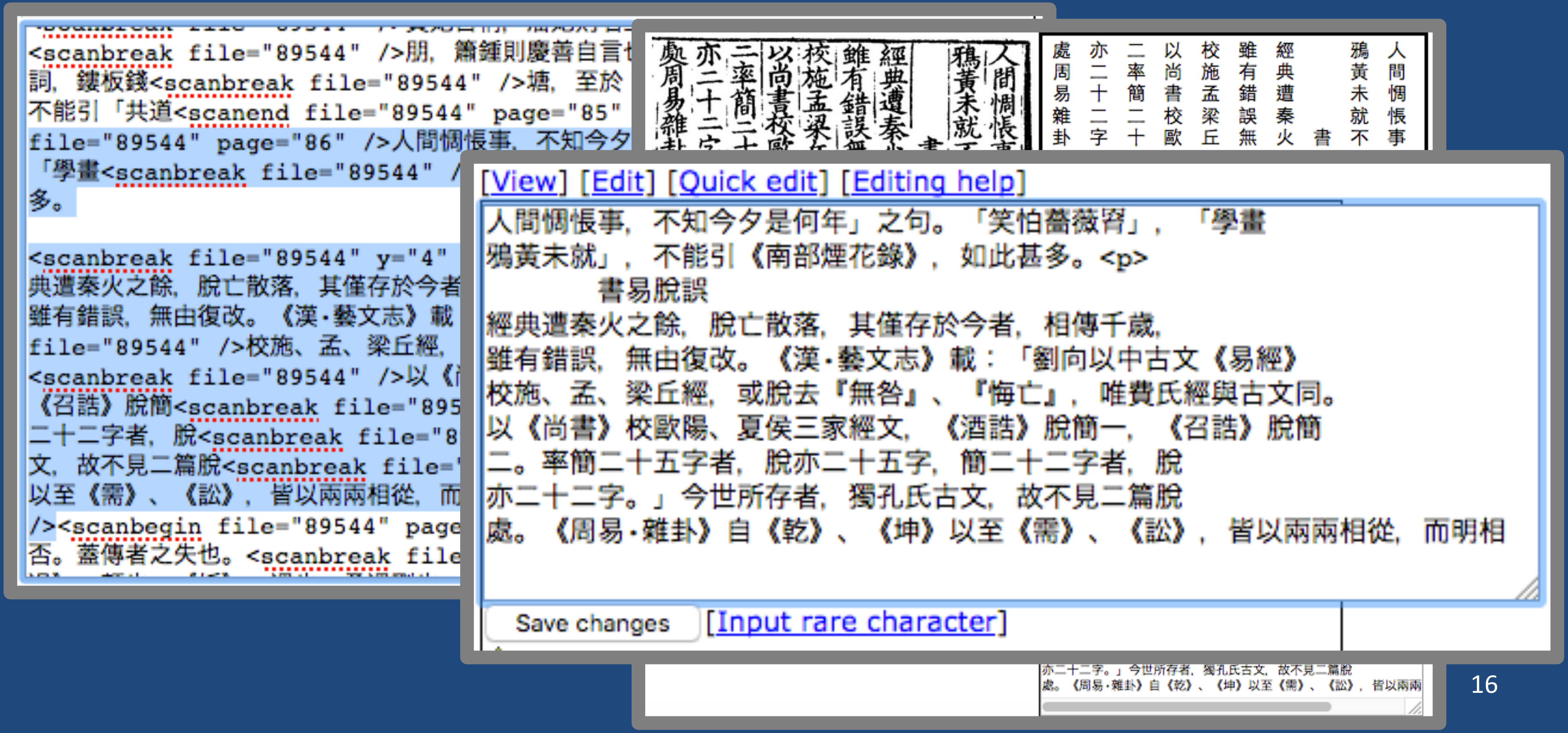
<scanbreak file="88886" />未嘗不嗚咽摧慟每忌日輒一日不食{{同上四百十三}}以儁為上郡王	<scanbreak file="88886" />未嘗不嗚咽摧慟每忌日輒三日不食{{同上四百十三}}以儁為上郡王



和為大司馬封梁
大夫呼延翼為大
延年年十五喪一
賊所掠延年追而
孤為叔父所養此
免之御覽四百一下傳有太宰
王洋御覽引誤作太宰王祥
每忌日輒三日不
王儁似誤字元英有幹
州路中糧匱密
逃免密乃將元直

Dynamic approach: challenges

- Typically need “human readable” serialization
  - Can be simplified using task-specific visual editors



```
<scanbreak file="89544" />朋，籥鍾則慶善自言
詞，鏤板錢<scanbreak file="89544" />塘，至於
不能引「共道<scanend file="89544" page="85"
file="89544" page="86" />人間惆悵事，不知今夕
「學畫<scanbreak file="89544" /
多。

<scanbreak file="89544" y="4"
典遭秦火之餘，脫亡散落，其僅存於今者
雖有錯誤，無由復改。《漢·藝文志》載
file="89544" />校施、孟、梁丘經，
<scanbreak file="89544" />以《
《召誥》脫簡<scanbreak file="895
二十二字者，脫<scanbreak file="8
文，故不見二篇脫<scanbreak file="
以至《需》、《訟》，皆以兩兩相從，而
/><scanbegin file="89544" page
否。蓋傳者之失也。<scanbreak file
```

人間惆悵事
鴉黃未就不
書
經典遭秦火
雖有錯誤無
校施孟梁丘
以尚書校歐
二率簡二十
亦二十二字
處周易雜卦

[View] [Edit] [Quick edit] [Editing help]

人間惆悵事，不知今夕是何年」之句。「笑怕薔薇罥」，「學畫
鴉黃未就」，不能引《南部煙花錄》，如此甚多。<p>
書易脫誤
經典遭秦火之餘，脫亡散落，其僅存於今者，相傳千歲，
雖有錯誤，無由復改。《漢·藝文志》載：「劉向以中古文《易經》
校施、孟、梁丘經，或脫去『無咎』、『悔亡』，唯費氏經與古文同。
以《尚書》校歐陽、夏侯三家經文，《酒誥》脫簡一，《召誥》脫簡
二。率簡二十五字者，脫亦二十五字，簡二十二字者，脫
亦二十二字。」今世所存者，獨孔氏古文，故不見二篇脫
處。《周易·雜卦》自《乾》、《坤》以至《需》、《訟》，皆以兩兩相從，而明相

Save changes [Input rare character]

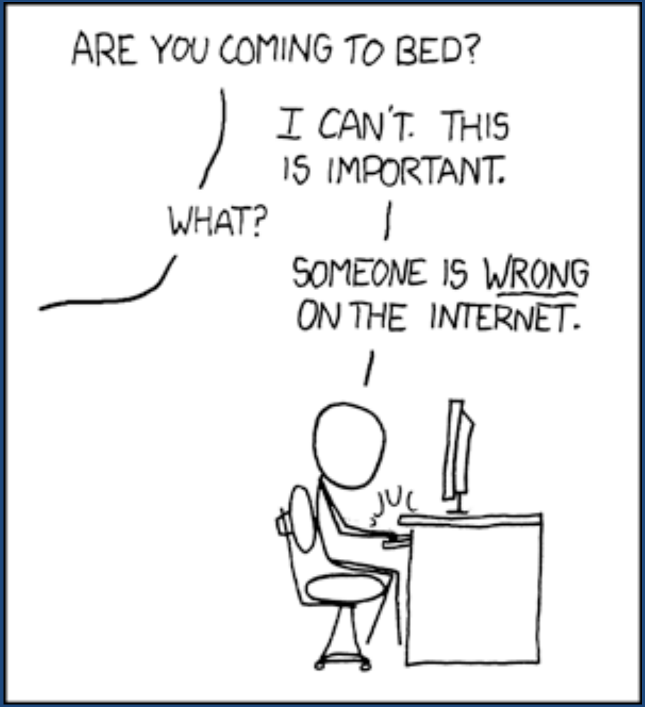
亦二十二字。」今世所存者，獨孔氏古文，故不見二篇脫
處。《周易·雜卦》自《乾》、《坤》以至《需》、《訟》，皆以兩兩

From database to platform

- Enable use as transcription & annotation tool
  - Image and textual data can be uploaded by users
  - OCR results can be corrected collaboratively
  - Future: semantic markup, translation, etc.
  - Full-text data can be exported via API
- Time and effort invested is preserved
  - Corrections *also* become part of public database
  - "Long tail" of non-mainstream material available

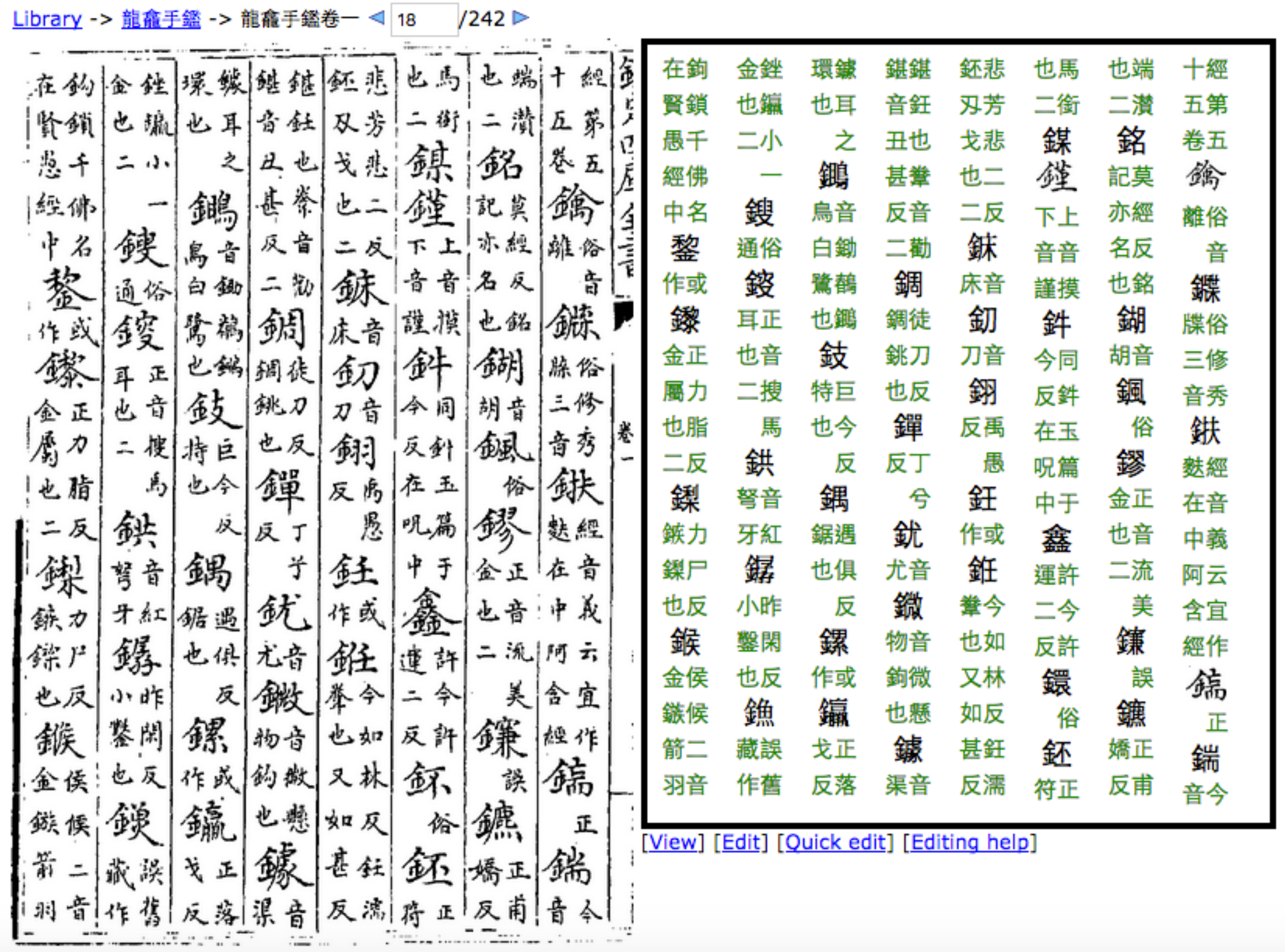
Psychological factors

- Creating *new* content vs correcting *incorrect* content
- Unpredictable response vs instant gratification
- Incremental improvement vs all-or-nothing approach


ARE YOU COMING TO BED?
I CAN'T. THIS IS IMPORTANT.
WHAT?
SOMEONE IS WRONG ON THE INTERNET.

https://xkcd.com/386/

OCR + Crowdsourcing



Library -> 龍龕手鑑 -> 龍龕手鑑卷一 ◀ 18 /242 ▶

十經五第卷五錀離俗音鐷牒俗三修音秀鈦麩經在音中義阿云含宜經作鎬正鍴音今

也端二濽銘記莫亦經名反也銘鍸胡音𨫍俗鏐金正也音二流美鐮誤鏕嬌正反甫

也馬二衙鍱鏠下上音音謹摸鉡今同反鉡在玉呪篇中于鑫運許二今反許鐶俗鉟符正

鉟悲夃芳戈悲也二二反鉥床音釖刀音𨥉反禹愚鈓作或鉟撆今也如又林如反甚鈓反濡

鍖鍖音鈓丑也甚撆反音二勸錭錭徒銚刀也反鐔反丁兮鈂尤音鐓物音鉤微也懸鐻渠音

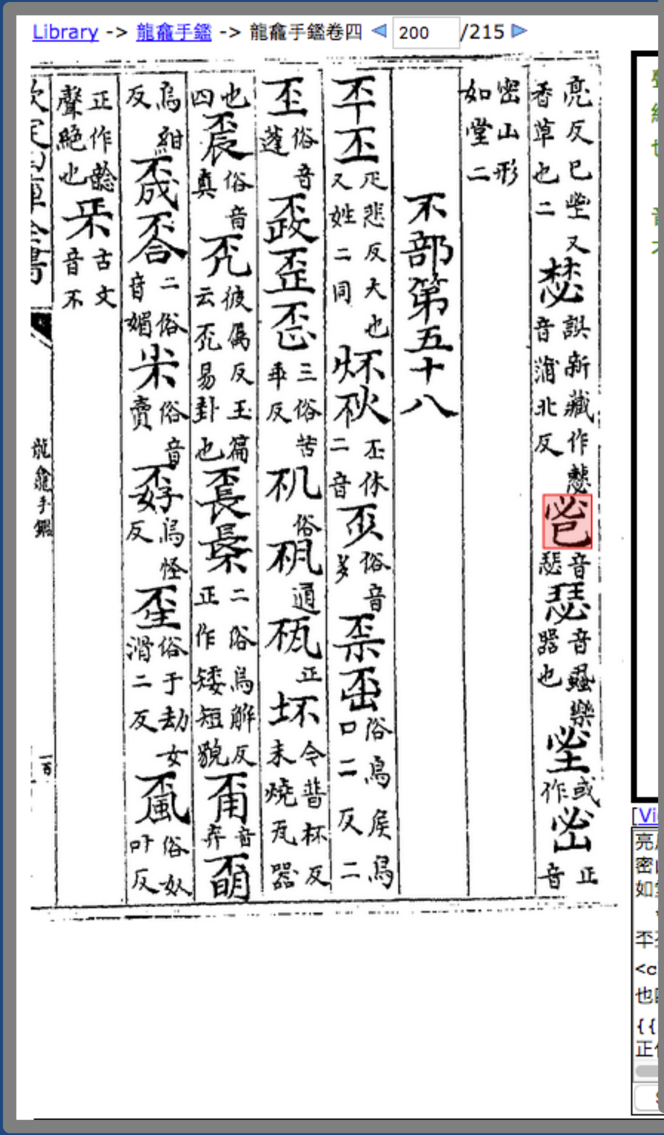
環鐻也耳之鎢烏音白鋤鷺鶴也鶮鈘特巨也今反鍝鋸遇也俱反鏍作或鎏戈正反落

金銼也鍽二小一鎪通俗鋑耳正也音二搜馬鉷弩音牙紅鎨小昨鑿閑也反𨧴藏誤作舊

在鉤賢鎖愚千經佛中名鍪作或鐩金正屬力也脂二反鍱鏉力鍱尸也反鍭金侯鏃候箭二羽音

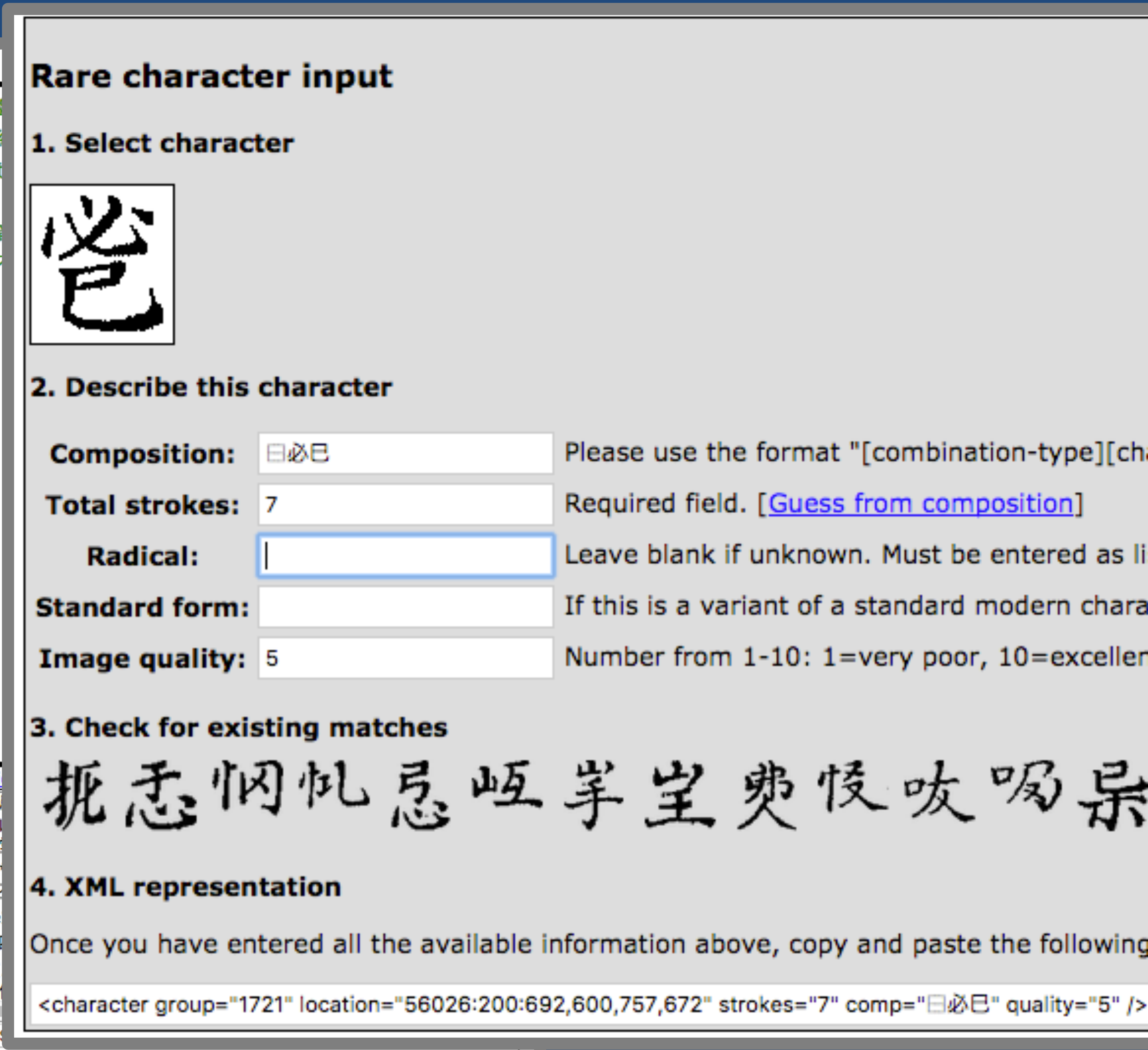
[View] [Edit] [Quick edit] [Editing help]

Task-specific visual editing tools



Library -> 龍龕手鑑 -> 龍龕手鑑卷四 ◀ 200 /215 ▶

不部第五十八



Rare character input

1. Select character

2. Describe this character

Composition:	⿱必巳	Please use the format "[combination-type][ch
Total strokes:	7	Required field. [Guess from composition]
Radical:		Leave blank if unknown. Must be entered as li
Standard form:		If this is a variant of a standard modern chara
Image quality:	5	Number from 1-10: 1=very poor, 10=exceller

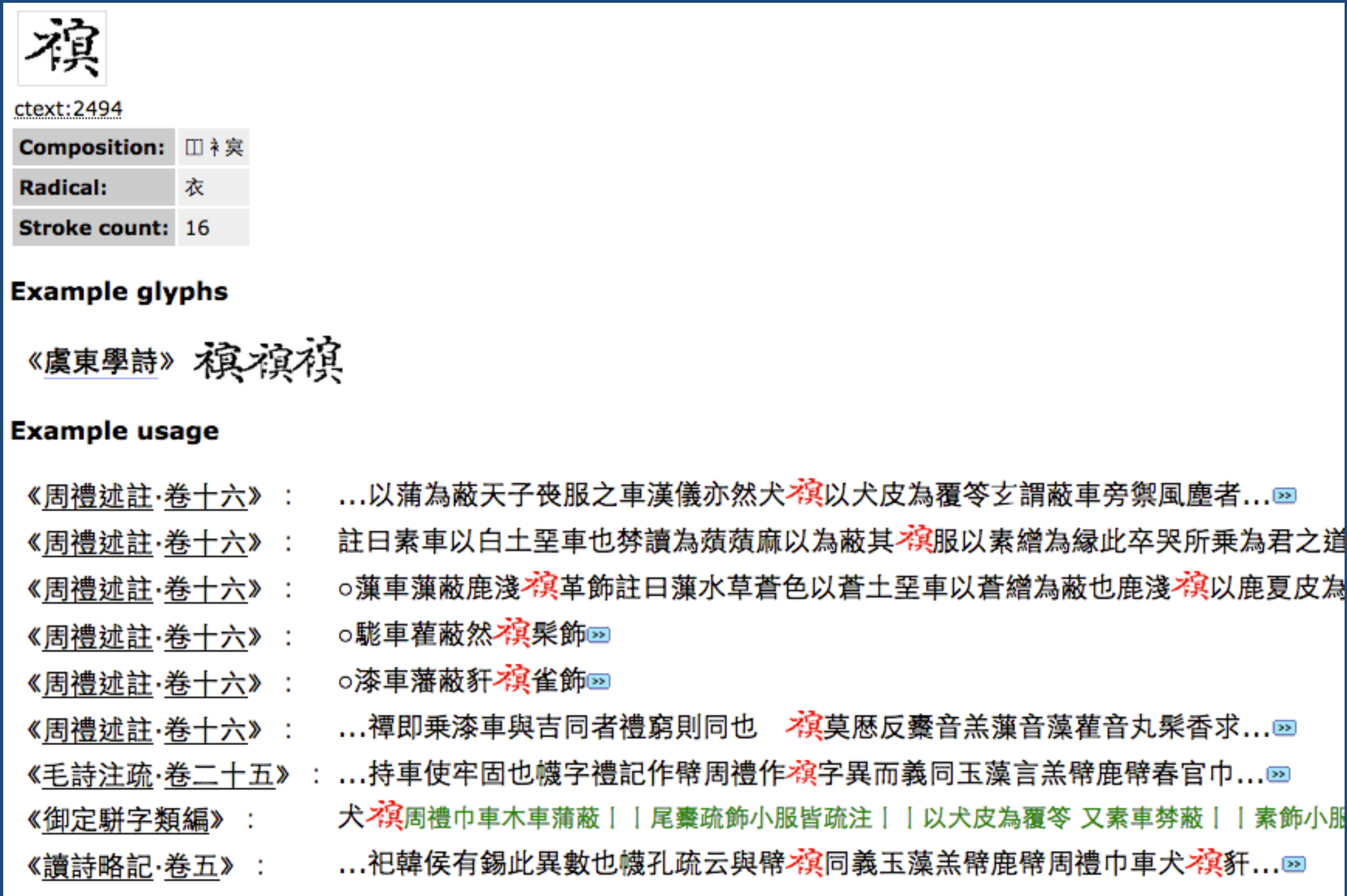
3. Check for existing matches

4. XML representation

Once you have entered all the available information above, copy and paste the following

<character group="1721" location="56026:200:692,600,757,672" strokes="7" comp="⿱必巳" quality="5" />

Collation of crowdsourced data



ctext:2494

Composition:	⿰衤寘
Radical:	衣
Stroke count:	16

Example glyphs

《虞東學詩》

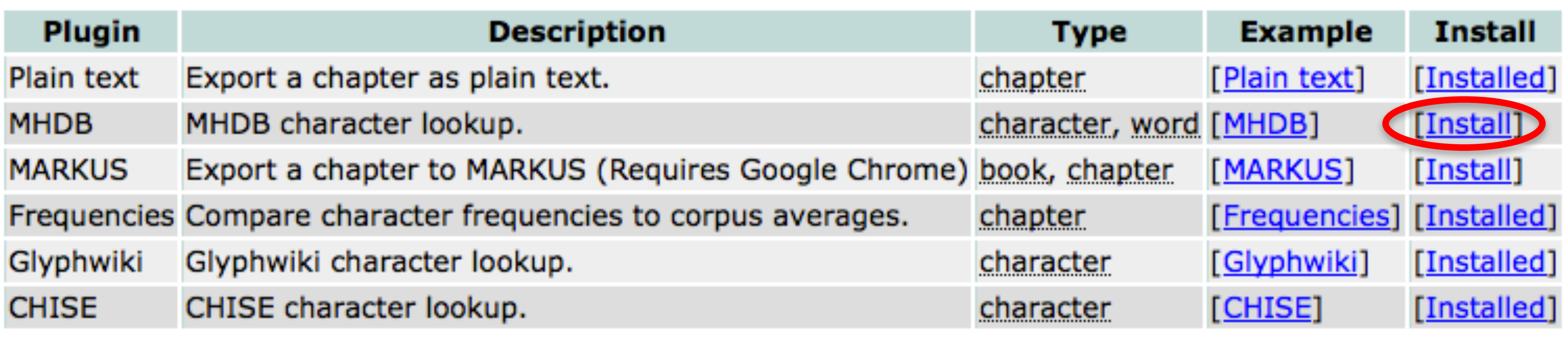
Example usage

《周禮述註·卷十六》	：	...以蒲為蔽天子喪服之車漢儀亦然犬⿰衤寘以犬皮為覆笭玄謂蔽車旁禦風塵者...»
《周禮述註·卷十六》	：	註曰素車以白土堊車也棼讀為薠薠麻以為蔽其⿰衤寘服以素繒為緣此卒哭所乘為君之道
《周禮述註·卷十六》	：	○藻車藻蔽鹿淺⿰衤寘革飾註曰藻水草蒼色以蒼土堊車以蒼繒為蔽也鹿淺⿰衤寘以鹿夏皮為
《周禮述註·卷十六》	：	○駹車雚蔽然⿰衤寘髤飾»
《周禮述註·卷十六》	：	○漆車藩蔽豻⿰衤寘雀飾»
《周禮述註·卷十六》	：	...禫即乘漆車與吉同者禮窮則同也　⿰衤寘莫歷反櫜音羔藻音藻雚音丸髤香求...»
《毛詩注疏·卷二十五》	：	...持車使牢固也幭字禮記作幦周禮作⿰衤寘字異而義同玉藻言羔幦鹿幦春官巾...»
《御定駢字類編》	：	犬⿰衤寘周禮巾車木車蒲蔽丨丨尾櫜疏飾小服皆疏注丨丨以犬皮為覆笭 又素車棼蔽丨丨素飾小
《讀詩略記·卷五》	：	...祀韓侯有錫此異數也幭孔疏云與幦⿰衤寘同義玉藻羔幦鹿幦周禮巾車犬⿰衤寘豻...»

CTP API Overview

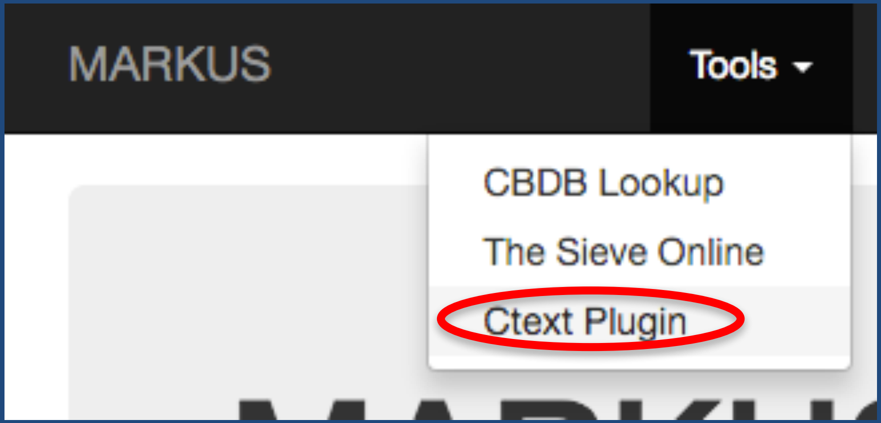
- CTP URNs
  - Unique identifier for each edition & chapter of a text
- JSON API
  - Extract machine-readable data from ctext.org
  - Textual data, metadata, etc.
- Plugin system
  - XML description of how to connect to another site
  - Allows users to connect ctext.org to other projects
  - User-definable, sharable, automatically updated

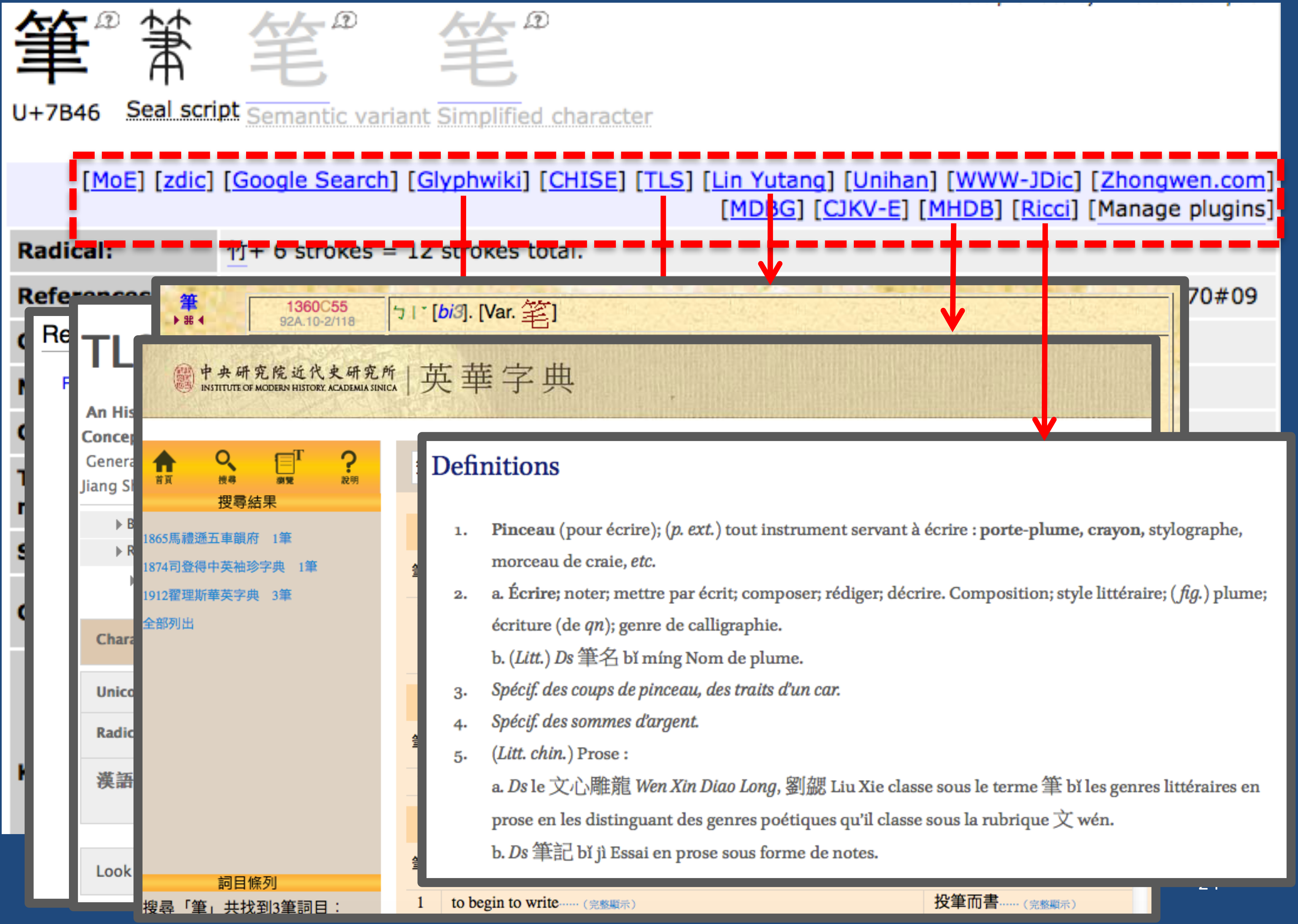
Plugins (User Perspective)



Plugin	Description	Type	Example	Install
Plain text	Export a chapter as plain text.	chapter	[Plain text]	[Installed]
MHDB	MHDB character lookup.	character, word	[MHDB]	[Install]
MARKUS	Export a chapter to MARKUS (Requires Google Chrome)	book, chapter	[MARKUS]	[Install]
Frequencies	Compare character frequencies to corpus averages.	chapter	[Frequencies]	[Installed]
Glyphwiki	Glyphwiki character lookup.	character	[Glyphwiki]	[Installed]
CHISE	CHISE character lookup.	character	[CHISE]	[Installed]

- Point-and-click installation within ctext.org
- Point-and-click installation from 3rd party site





筆 𥬙 笔 笔

U+7B46 Seal script Semantic variant Simplified character

[MoE] [zdic] [Google Search] [Glyphwiki] [CHISE] [TLS] [Lin Yutang] [Unihan] [WWW-JDic] [Zhongwen.com] [MDBG] [CJKV-E] [MHDB] [Ricci] [Manage plugins]

Radical: 竹 + 6 strokes = 12 strokes total.

筆 ▸⌘◂ | 1360C55 92A.10-2/118 | ㄅㄧˇ [bi3]. [Var. 笔]

中央研究院近代史研究所 INSTITUTE OF MODERN HISTORY, ACADEMIA SINICA | 英華字典

1865馬禮遜五車韻府 1筆

1874司登得中英袖珍字典 1筆

1912翟理斯華英字典 3筆

全部列出

詞目條列

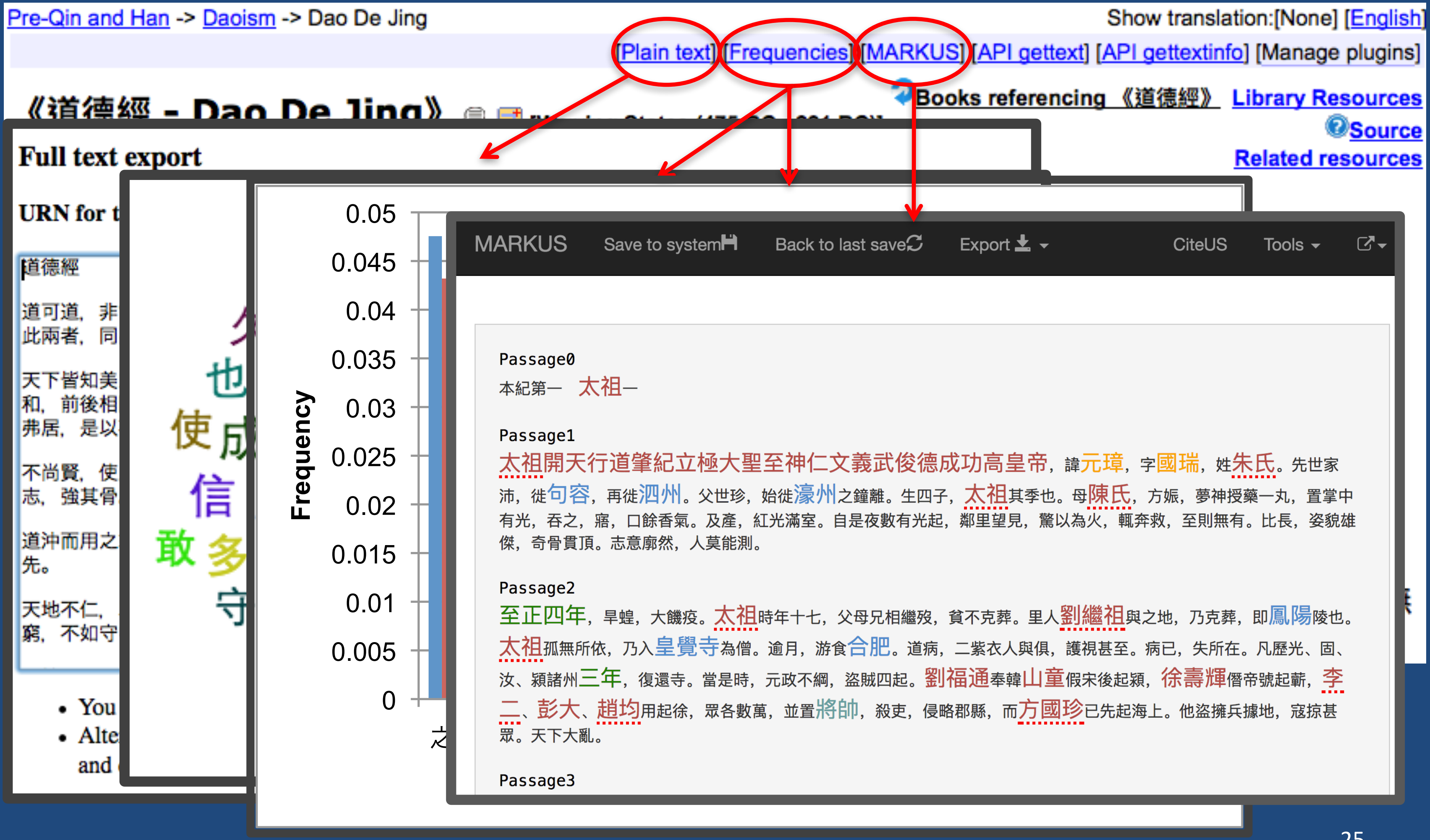
搜尋「筆」共找到3筆詞目：

1 to begin to write……（完整顯示） 投筆而書……（完整顯示）

Definitions

1. **Pinceau** (pour écrire); (*p. ext.*) tout instrument servant à écrire : **porte-plume, crayon,** stylographe, morceau de craie, *etc.*
2. a. **Écrire**; noter; mettre par écrit; composer; rédiger; décrire. Composition; style littéraire; (*fig.*) plume; écriture (de *qn*); genre de calligraphie.
   b. (*Litt.*) *Ds* 筆名 bǐ míng Nom de plume.
3. *Spécif. des coups de pinceau, des traits d'un car.*
4. *Spécif. des sommes d'argent.*
5. (*Litt. chin.*) Prose :
   a. *Ds* le 文心雕龍 *Wen Xin Diao Long*, 劉勰 Liu Xie classe sous le terme 筆 bǐ les genres littéraires en prose en les distinguant des genres poétiques qu'il classe sous la rubrique 文 wén.
   b. *Ds* 筆記 bǐ jì Essai en prose sous forme de notes.

API and Plugins



Full text export

URN for t

道德經

道可道，非
此兩者，同

天下皆知美
和，前後相
弗居，是以

不尚賢，使
志，強其骨

道沖而用之
先。

天地不仁，
窮，不如守

- You
- Alte and

Passage0

本紀第一　太祖一

Passage1

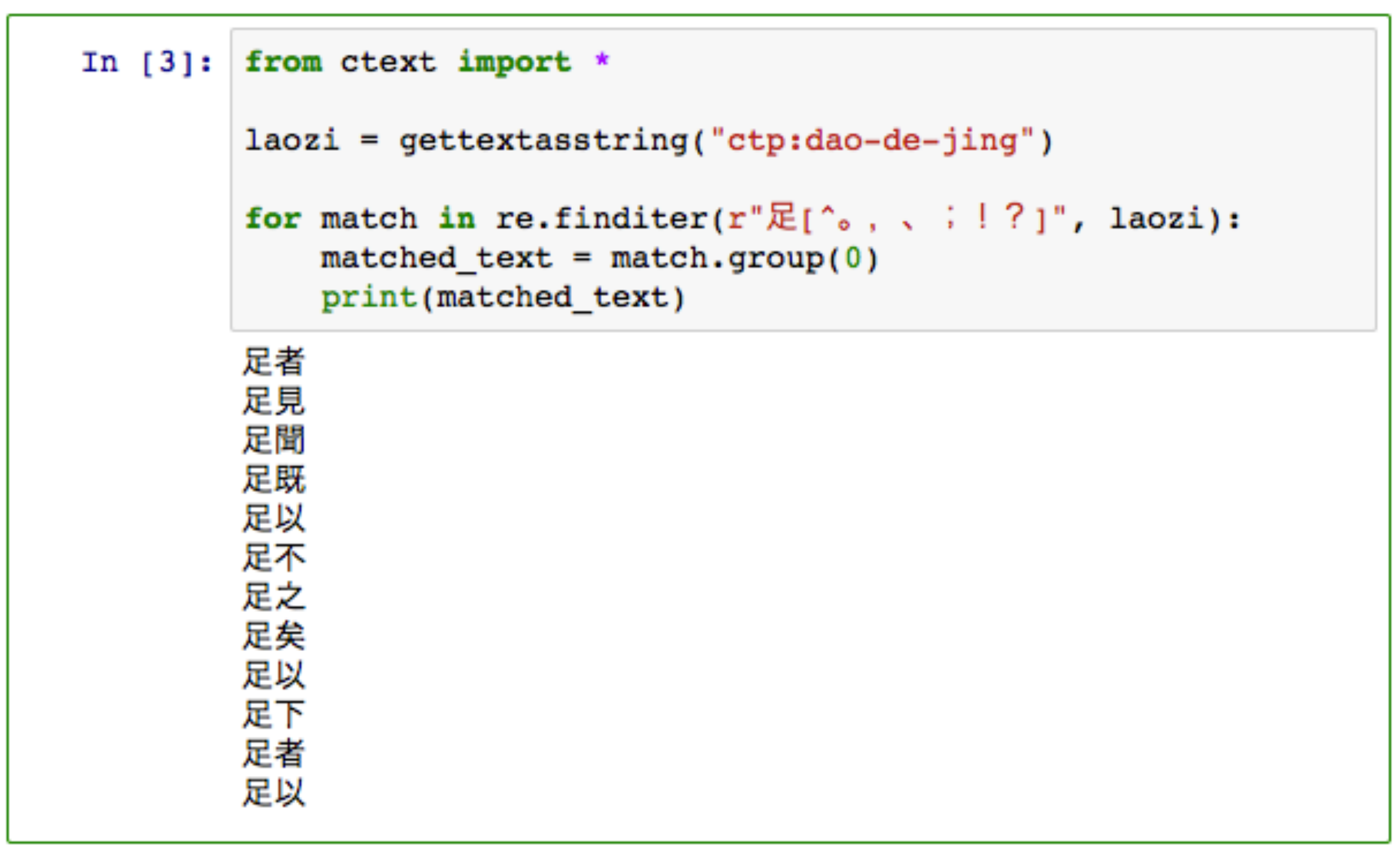
太祖開天行道肇紀立極大聖至神仁文義武俊德成功高皇帝，諱元璋，字國瑞，姓朱氏。先世家沛，徙句容，再徙泗州。父世珍，始徙濠州之鐘離。生四子，太祖其季也。母陳氏，方娠，夢神授藥一丸，置掌中有光，吞之，寤，口餘香氣。及產，紅光滿室。自是夜數有光起，鄰里望見，驚以為火，輒奔救，至則無有。比長，姿貌雄傑，奇骨貫頂。志意廓然，人莫能測。

Passage2

至正四年，旱蝗，大饑疫。太祖時年十七，父母兄相繼歿，貧不克葬。里人劉繼祖與之地，乃克葬，即鳳陽陵也。太祖孤無所依，乃入皇覺寺為僧。逾月，游食合肥。道病，二紫衣人與俱，護視甚至。病已，失所在。凡歷光、固、汝、潁諸州三年，復還寺。當是時，元政不綱，盜賊四起。劉福通奉韓山童假宋後起潁，徐壽輝僭帝號起蘄，李二、彭大、趙均用起徐，眾各數萬，並置將帥，殺吏，侵略郡縣，而方國珍已先起海上。他盜擁兵據地，寇掠甚眾。天下大亂。

Passage3

APIs in Teaching



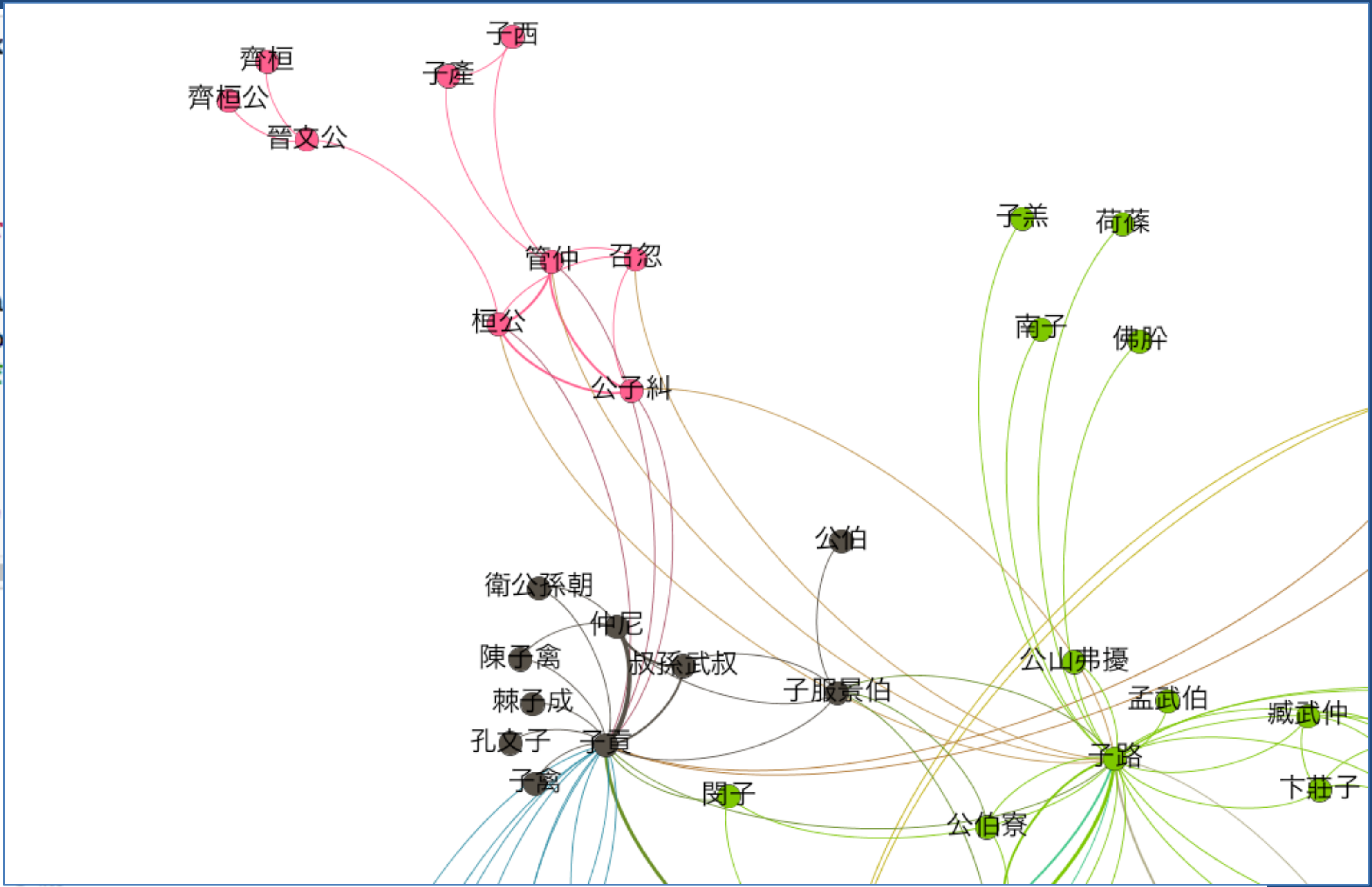
```
In [3]: from ctext import *

        laozi = gettextasstring("ctp:dao-de-jing")

        for match in re.finditer(r"足[^。，、；！？]", laozi):
            matched_text = match.group(0)
            print(matched_text)
```

```
足者
足見
足聞
足既
足以
足不
足之
足矣
足以
足下
足者
足以
```

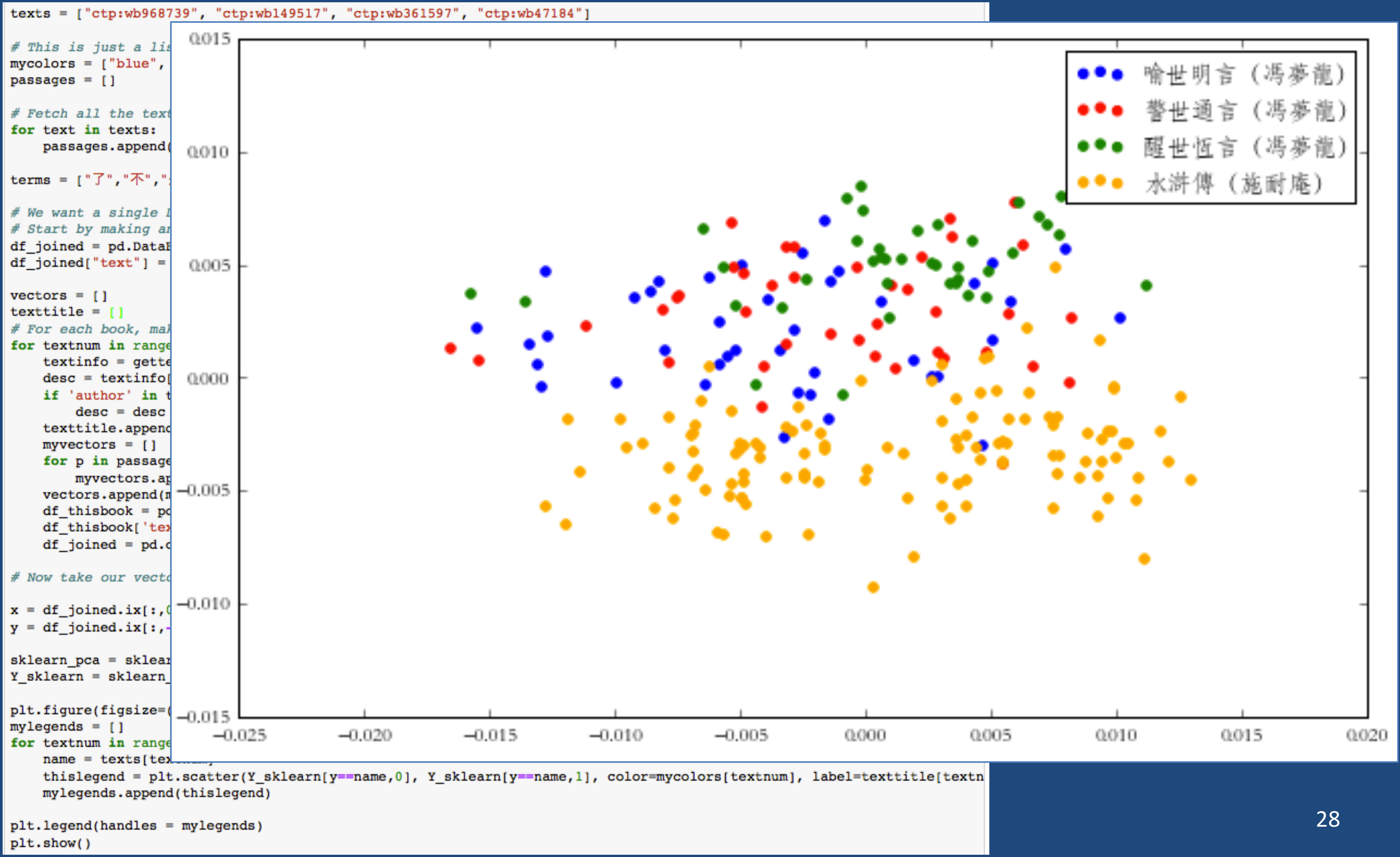
APIs in Teaching

```
from ctex

passages
people =

print("gr

for passa
    for p
        f



print("}"
```

```
graph {
  樊遲 --
  顏淵 --
  顏淵 --
  季路 --
  顏淵 --
  子貢 --
  子貢 --
  子貢 -- 顏淵
```



APIs in Teaching



```
texts = ["ctp:wb968739", "ctp:wb149517", "ctp:wb361597", "ctp:wb47184"]

# This is just a lis
mycolors = ["blue",
passages = []

# Fetch all the text
for text in texts:
    passages.append(

terms = ["了","不","

# We want a single
# Start by making an
df_joined = pd.Data
df_joined["text"] =

vectors = []
texttitle = []
# For each book, mak
for textnum in range
    textinfo = gette
    desc = textinfo[
    if 'author' in t
        desc = desc
    texttitle.append
    myvectors = []
    for p in passage
        myvectors.ap
    vectors.append(m
    df_thisbook = pd
    df_thisbook['tex
    df_joined = pd.c

# Now take our vecto

x = df_joined.ix[:,0
y = df_joined.ix[:,-

sklearn_pca = sklear
Y_sklearn = sklearn_

plt.figure(figsize=(
mylegends = []
for textnum in range
    name = texts[te
    thislegend = plt.scatter(Y_sklearn[y==name,0], Y_sklearn[y==name,1], color=mycolors[textnum], label=texttitle[textn
    mylegends.append(thislegend)

plt.legend(handles = mylegends)
plt.show()
```

Conclusion

- OCR & crowdsourcing allow access to long tail
  - Obscure texts many of which are untranscribed
  - Infrastructure helps users help themselves
    - Side effect: their efforts benefit everyone else too
- API access
  - Machine-friendly interface enables new use cases
  - Allows use in many workflows
  - Streamline DH teaching and research
    - Avoids dilemma of toy examples vs data wrangling

More information:

ctext.org

Practical introduction:

dsturgeon.net/ctext