視覺化古籍校勘平台
——人文與科技協作之嘗試

2014.6.5 古籍全文資料庫的回顧與展望工作坊

Yap Cheah Shen

Agenda

視覺化校勘平台功能概括

古書的引用--以康熙字典為例

數位化引用--互文超連結

視覺化校勘平台使用之技術

演示

曾參與之計劃

1991 開始參與佛典數位化

1996 佛光大辭典光碟版

1999 印順法師佛學著作集 www.yinshun.org.tw

2006 巴利大藏經線上搜尋 www.tipitaka.org

2010 開放康熙字典 kangxi.adcs.org.tw

2013 藏文大藏經計劃

視覺化古籍校勘平台功能概括

一）校對：輸入文字與原書圖檔比對

二）句讀、分段標記。

三）對勘：對照其他版本的圖檔及文字。

四）引文溯源。

五）產生集注。

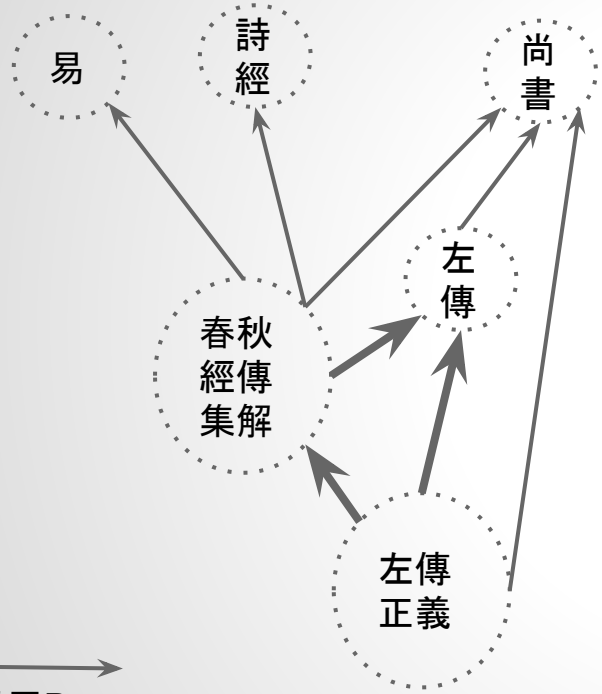
古書引用的種類及建立方法

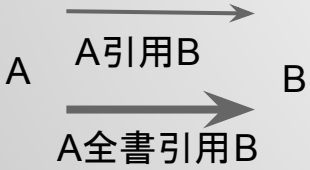
明引：標明出處。《左傳．莊九年》：公喪戎路，傳乘而歸。

暗引：出處不明。古德曰、先賢曰、經曰。

檃括：將前人詩句加以裁剪，或增字、減字、或改字。
蘇軾《哨遍．檃括＜歸去來辭＞》

引用網路





一）引用所構成非循環(acyclic)網路，即從任何一個節點出發，不可能回到自身。

二）違反以上原則，可能是後期夾入之文字。

建立引用的方法

「全文檢索」可檢出明引、暗引之出處。

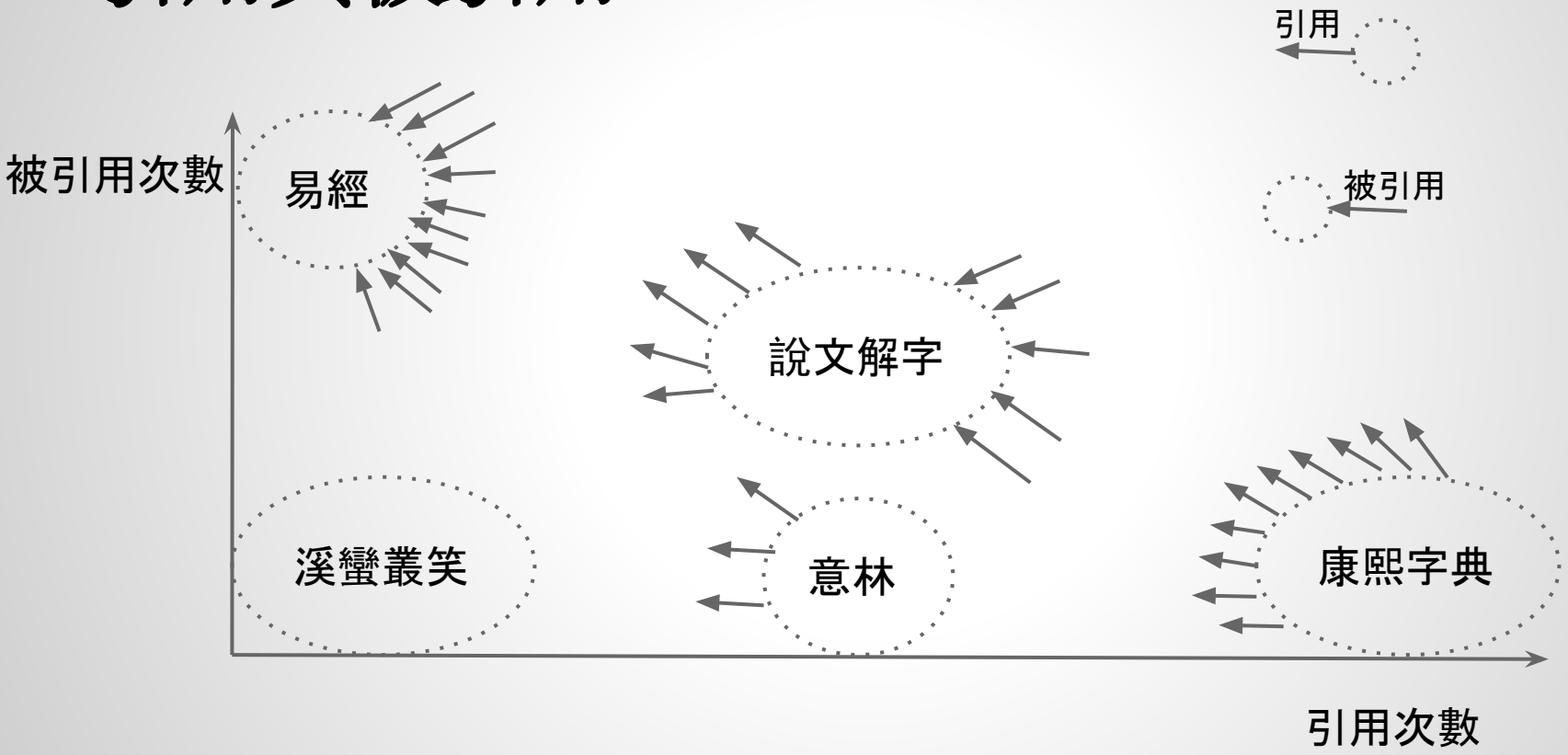
檃括則必須先以「近似檢索」找出可能出處，再由內容專家確認。

被引用之重要性

Google PageRank：由某一網頁被其他網頁連結的次數，決定該網頁之重要性。
http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf
http://infolab.stanford.edu/~backrub/google.html

經由古書的被引用情況，例如：被那些書引用、跨越的時代、數量等，作為評量重要性的工具。

引用與被引用



基於引用方式的內容分類

	引用多書多次 (工具類)	**引用某書多次 (注疏類)**	**很少引用 (原創類)**
被很多書引用多次	《說文解字》《太平御覽》	《左傳》《經典釋文》	《易經》《道德經》《詩經 》
被某些書引用多次	專門書籍《通志堂經解》《說文解字詁林》《廿二史考異》《十七史商榷》	《水經注》《輿地紀勝》	某一領域之經典《內經》《孫子兵法》東京夢華錄》
很少被引用	《意林》《紺珠集》	《晁氏客語》《芥隱筆記》	《溪蠻叢笑》《海槎餘錄》

康熙字典引用分析

開放古籍協會康熙字典新式標點	舉例	引用書目本數	引用次數	百分比	百分比(不含>1000次字書)
所有引用書目		2015	187,413	100%	
引用超過1000次	說文、易經	25	163,271	87.12%	
>1000次字書	說文、集韻	15	138,439	73.86%	
>1000次字書以外所有書		2000	48,974	26.13%	100%
>1000次非字書	見下頁	10	24,832	13.24%	50.70%
<1000次		1990	24,142	12.88%	49.30%

康熙字典引用超過1000次書目(不含字書)

書名	總引用次數 24,832	總字數	每百字引用次數	推測成書年代
詩經	4597	32,962	13.946	BC1000~BC400
前漢書	3994	804,156	0.496	AD 89~AD 130
禮記	3143	101,632	3.092	BC403~BC221
史記	2691	580,171	0.463	BC104~BC90
左傳	2589	198,180	1.306	BC400~BC296
周禮	2419	96,520	2.506	戰國中期(徐復觀)
尚書	1963	25,946	7.565	戰國時期
易經	1219	18,437	6.612	西周初/西周末
後漢書	1193	1,384,633	0.086	AD 445
莊子	1024	67,495	1.517	戰國末年

引用與被引用的數位表達方式

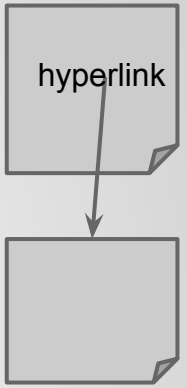
超連結:起點為任意文字，目標為檔案(節點)

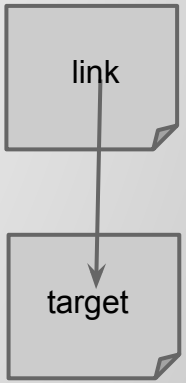
HTML的<a>或TEI 的<link><ptr><ref>

引用:起點和目標皆是任意文字

HTML 和TEI皆無法描述任意文字的目標。

需要設計新的機制來表達「互文超連結」。





數位化引用(互文超連結)所需之技術

一)跨版本任意段落文獻定址

資料結構 citation = {source:[start,len], target: [start,len]}

二)古籍校勘及標注平台

讓人文學者以視覺化方式輸入對文本之理解。例如:建立引用連結, 內容類型, 詞組類型。

跨版本任意段落文獻定址

郵遞地址 (1874 Universal Postal Union)

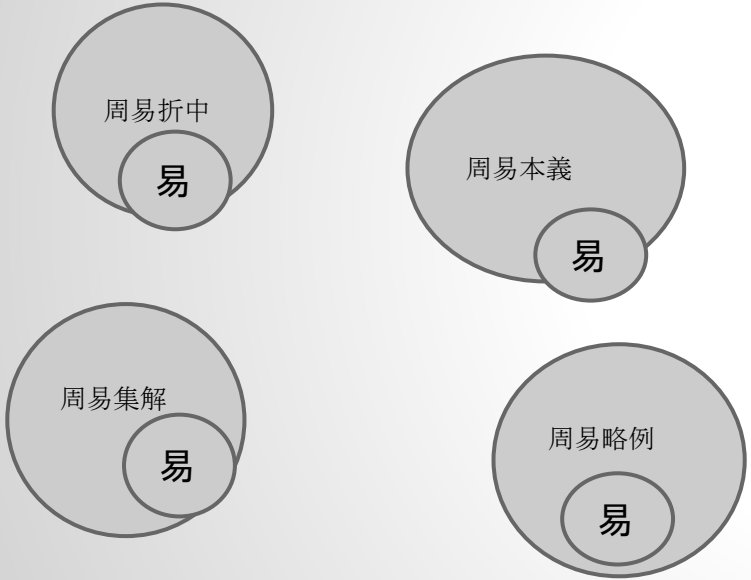
IP (Internet protocol) 位址 (RFC791 1981)

古書的Unified Book ID(ISBN),Author ID, per-character ID

e.g. 傳世本Bible(跨版本、跨語言) vs

Chinese Tripitaka(冊頁碼不能跨版本、語言)

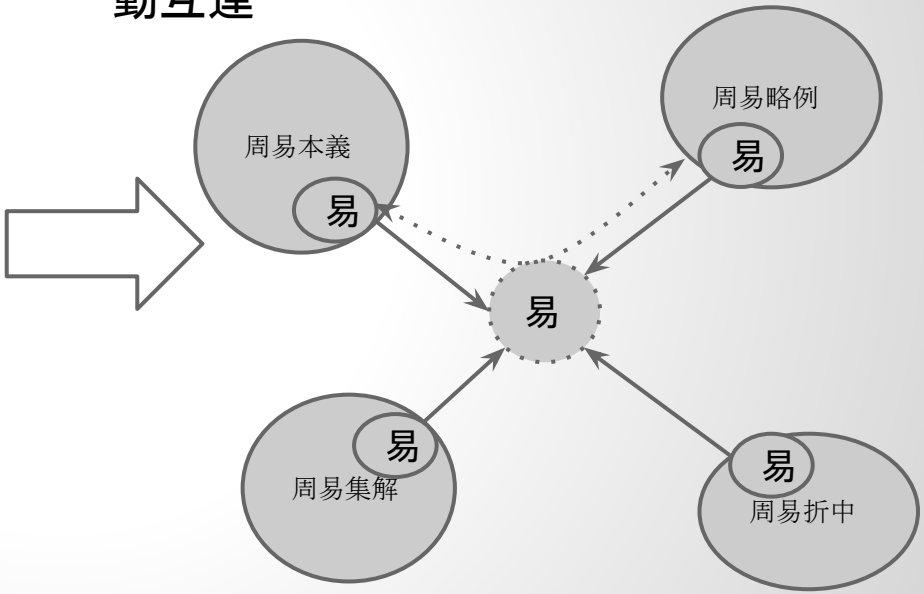
從被引用產生集注



注疏嵌入全部或部份原文。

不易查找某一段經文的其他解釋。

雙箭頭虛線：注疏經由統一版本的《易》自動互連



將引用標記之後，電腦可自動整合同段經文的不同解釋。

視覺化校勘平台介紹

一）簡單易用。

二）網路應用程式或離線應用程式。

三）使用Github 等現有技術架構。

四）自由軟體，可依需求客制化。

系統模組

一）全文及近似句搜尋。

二）視覺化標記，不假設使用者懂XML。

三）使用者可以選擇任意兩段文字，建立連結。

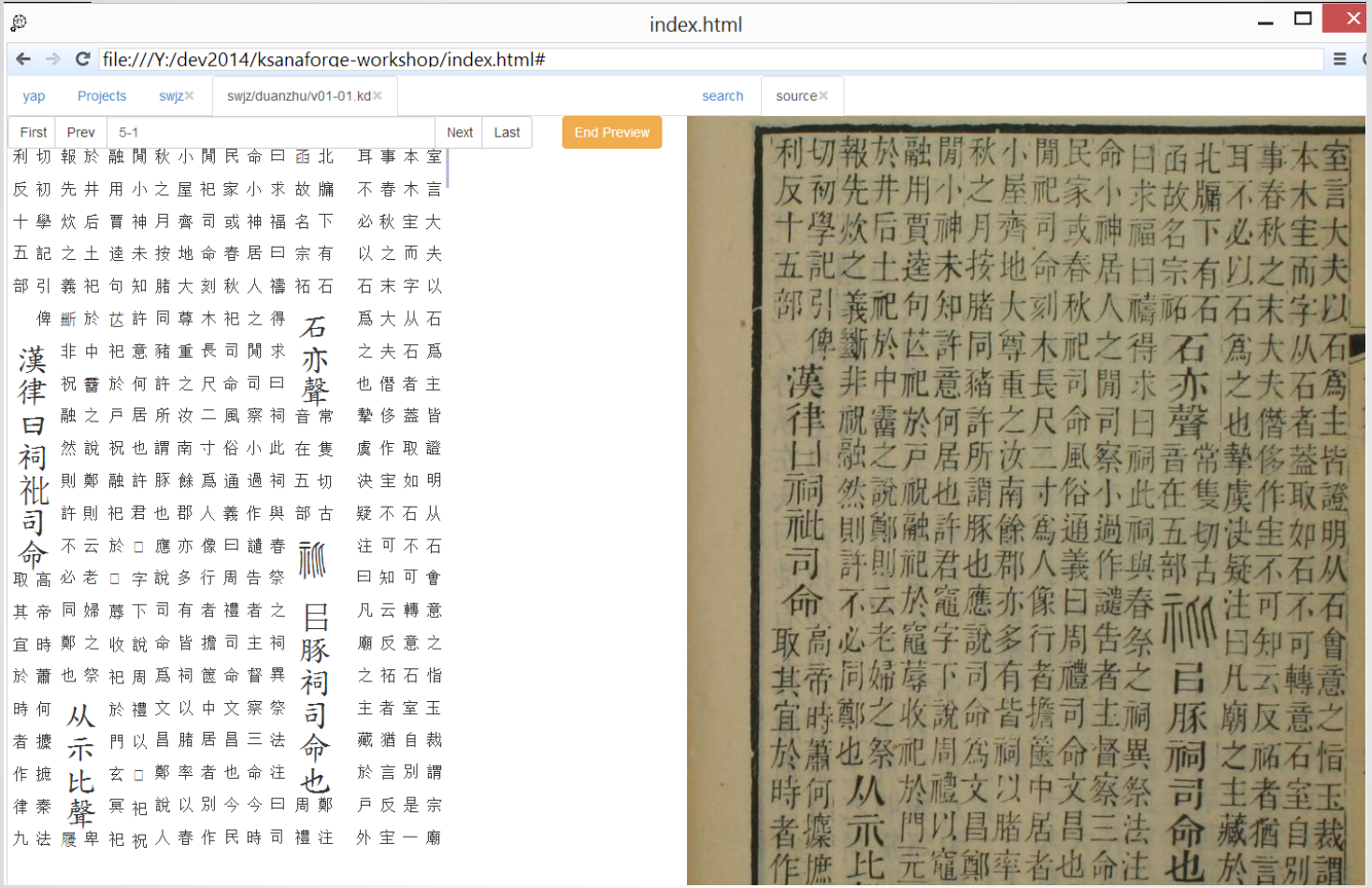
四）跨版本連結。對A版本的引用，也可以自動連到同書的B版本。

採用之技術

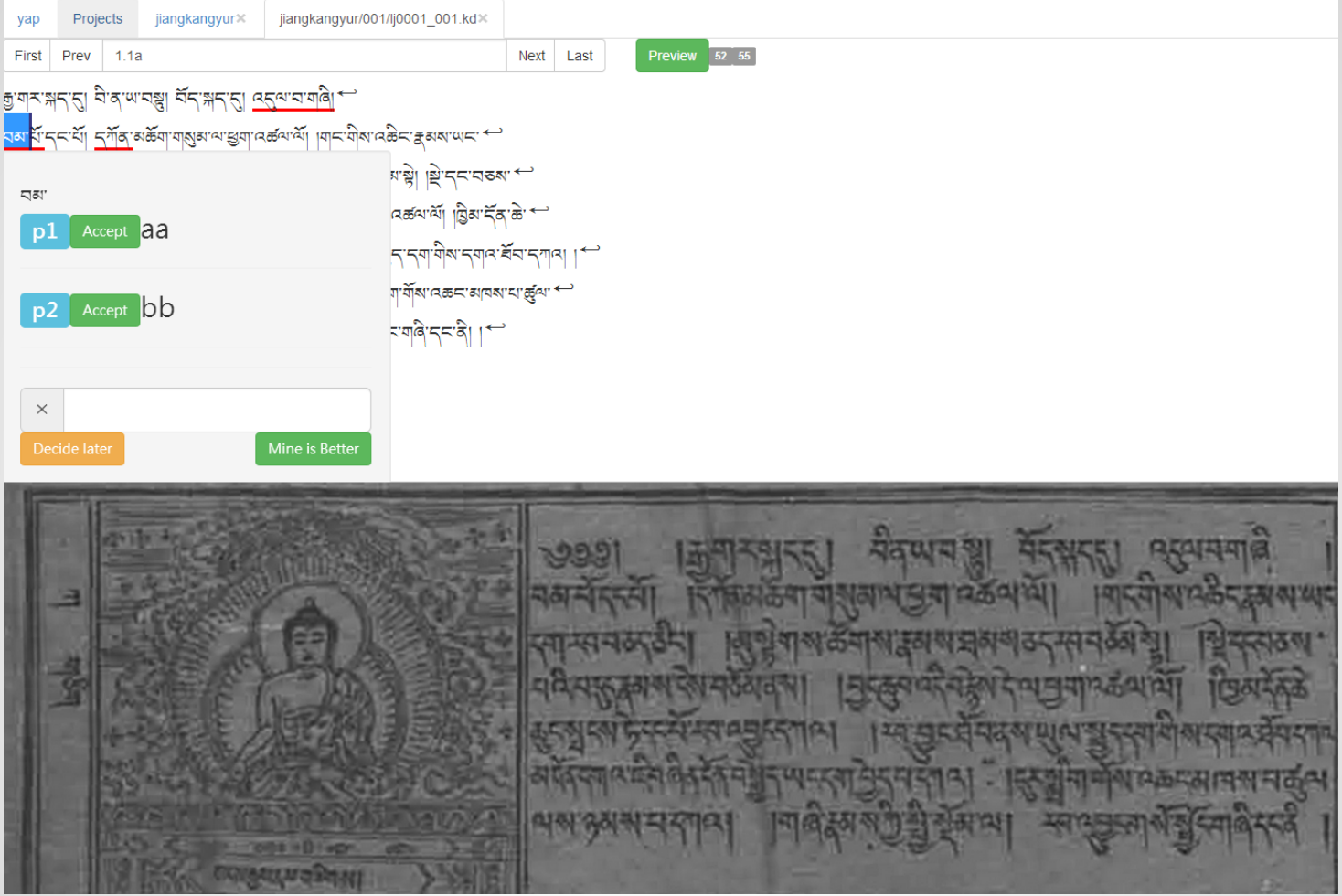
一）HTML5+CSS+Javascript

二）Node.js (Windows, Mac OS X, Linux)

三）Open Data Format (JSON)



使用css3 vertical layout 仿古書夾注版型



圖文比對，視覺化輸入校勘。

近似句搜尋

十三經溯源

《唐韻》：陟弓切；《集韻》、《韻會》、《正韻》：陟隆切，
《書・大禹謨》：允執厥中。《周禮・地官・大司徒》：以五禮
教之中。《左傳・成十三年》：劉子曰：民受天地之中以生。
又，《左傳・文元年》：舉正於中，民則不惑。〔註〕舉中氣也
又，司中，星名，在太微垣。《周禮・春官・大宗伯》：以槱燎
司命、飌師、雨師。
又，《前漢・律歷志》：春爲陽中，萬物以生；秋爲陰中，萬物
又，中央，四方之中也。《書・召誥》：王來紹上帝，自服于土
〔註〕洛爲天地之中。班固〈東都賦〉：宅中圓大。
又，正也。《禮・儒行》：儒有衣冠中。《周禮・春官・司刺》

列出連結

連結數27筆

《書・大禹謨》	允執厥中
100% 虞書.大禹謨 原文	
《周禮・地官・大司徒》	以五禮防民僞而教之中
44.44% 大司徒 原文	
《左傳・成十三年》	劉子曰：民受天地之中以生
88.89% 左傳.成公.傳十三年 原文	

說明 左傳.成公.傳十三年

賜・請先使・王以行人之
禮・禮焉・孟獻子從・王以為介・而重賄之・公及諸
伐秦・成子受脤于社・不敬・劉
子曰・吾聞之・民受天地之中以生・所謂命也・是以
養之以福・不能者敗以
取禍・是故君子勤禮・小人盡力・勤禮莫如致敬・盡
之大事・在祀與戎・祀有執
膰・戎有受脤・神之大節也・今成子惰弃其命矣・其
絕秦・曰・昔逮我獻公・及穆
公相好・戮力同心・申之以盟誓・重之以昏姻・天禍
公即世・穆公不忘舊德・俾我惠公・用能奉祀于晉・
心・用集我文公・是穆之成也・文公躬擐甲冑・跋履
川・踰越險阻・征東之諸侯・虞夏商周之胤・而朝諸
場・我文公帥諸侯及秦圍
鄭・秦大夫不詢于我寡君・擅及鄭盟・諸侯疾之・將
克還無害・則是我有大造
于西也・無祿・文公即世・穆為不弔・蔑死我君・寡
城・殄滅我費滑・散離我兄弟・
撓亂我同盟・傾覆我國家・我襄公未忘君之舊勳・而
于穆公・穆公弗聽・而
即楚謀我・天誘其衷・成王隕命・穆公是以不克逞志
自出・又欲闕翦我公室・
傾覆我社稷・帥我蝥賊・以來蕩搖我邊疆・我是以有
涑川・俘我王官・翦我羈
馬・我是以有河曲之戰・東道之不通・則是康公絕我
望曰・庶撫我乎・君亦不

開放康熙字典 福 57:8

《唐韻》、《集韻》、《韻會》方六切，膚平聲——祐也、休也、善也、祥也。《禮．祭統》：福者，備也。《易．謙卦》：鬼神害盈而福謙。《書．洪範》：嚮用五福。↩

又，《釋名》：福，富也——其中多品如富者也

又，祭祀胙肉曰福。《周禮．天官．膳夫》：祭祀之致福者，受而膳之。《穀梁傳．僖十年》：祠致福於君。↩

又，福，猶同也。張衡〈西京賦〉：仰福帝居，陽曜陰藏。〔薛註〕言今長安宮，上與「五帝所居之太微宮，陽時則見、陰時則藏」同法也。↩

又，州名。秦閩中郡，陳立閩州，唐改福州。↩

十三經 謙 48:8

䷎謙．亨．君子有終．彖曰．謙亨．天道下濟而光明．地道卑而上行．天道虧盈而益謙．地道變盈而流謙．鬼神害盈而福謙．人道惡盈而好謙．謙尊而光．卑而不可踰．君子之終也．象曰．地中有山．謙．君子以裒多益寡．稱物平施．初六．謙謙君子．用涉大川．吉．象曰．謙謙君子．卑以自牧也．六二．鳴謙．貞吉．象曰．鳴謙貞吉．中心得也．九三．勞謙君子．有終吉．象曰．勞謙君子．萬民服也．六四．无不利撝謙．象曰．无不利撝謙．不違則也．

允許雙選取區，任意段落互連。

Live Demo

請多指教

yapcheahshen@gmail.com
http://www.ksana.tw

文字資料庫的三個階段

階段	數位化的標的	資訊技術	人文知識
Morphology	內容本身	輸入法, OCR 編輯器，缺字處理	識字
Metadata	內容之外的情境	資料庫 XML標記語言 檢索工具	編目分類
Meaning	內容之間的情境	相似檢索 視覺化標記工具	文獻學家 專精內容者

內容理解的前置工作:校勘

文獻學知識:版本、目錄、考證、輯佚、校勘。

校勘:一日存真,二日校異,三日訂訛。

電腦能幫忙的事:

校異、訂訛:保留各種版本之差異,供人工比對、挑選。

存真:保留所有原書圖片,同時呈現。

輯佚:DNA Sequencing,自動拼圖。

考證:找出資料之出處。引用。