理论与技术

古 籍 数 字 化 实 践

朱岩

（北京书同文数字化技术有限公司）

中文古籍数字化在技术方面已经成熟。主要体现在一批最有代表性、大部头的《四部丛刊》、《四库全书》等典籍，已成功地实现了数字化。一些学者反映，过去从事一项课题研究，常常要花上三、四个月时间搜集资料，而且尚无查全把握。现在，利用数字化的《四部丛刊》、《四库全书》检索可在瞬间完成，再花一点时间确认、拷贝，马上就可投入新的研究及其论文撰写，科研工作效率显著提高，而且对古籍内容挖掘的深度和广度也是过去手工办法无法比拟。本文简要介绍《四部》、《四库》等古籍数字化方面的经验。

一、 合理目标定位

古籍数字化能否成功，合理的目标定位至关重要。

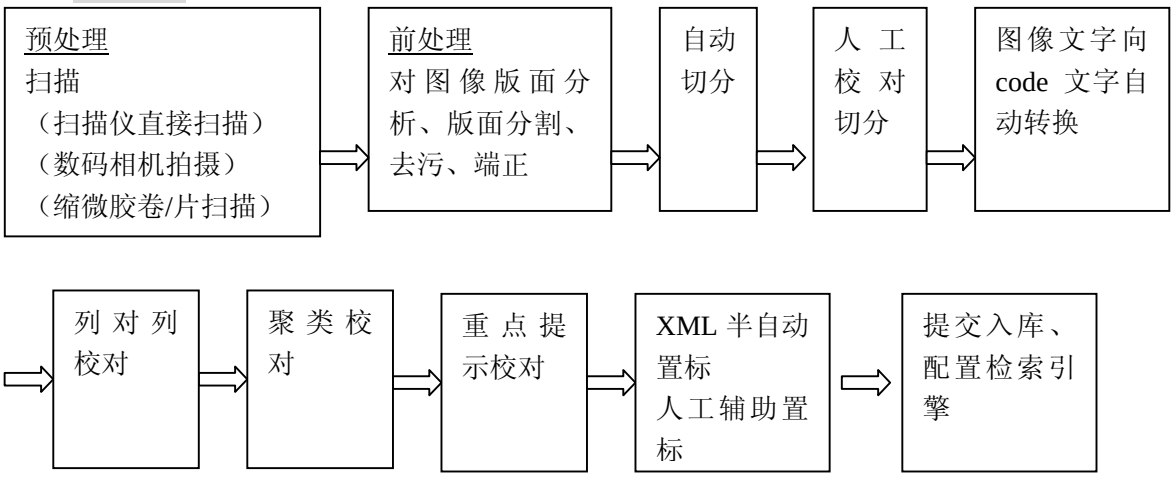
对于以文字为主的中国古籍来说，它的数字化绝不是纸张载体版本的翻版。扫描是必要的，但扫描在很多情况下只是数字化的预处理。根据我们的实践，扫描仅占数字化工程的2%。

把古籍的内容数字化并使之与多种有效的检索、处理工具完美结合，奉献给读者知识宝库和卓有成效的研究手段，使学者多出成果，快出成果，这才是古籍数字化的目标。

还有一点不可忽视的是，古籍数字化后的文字，差错率应达到出版界规定。

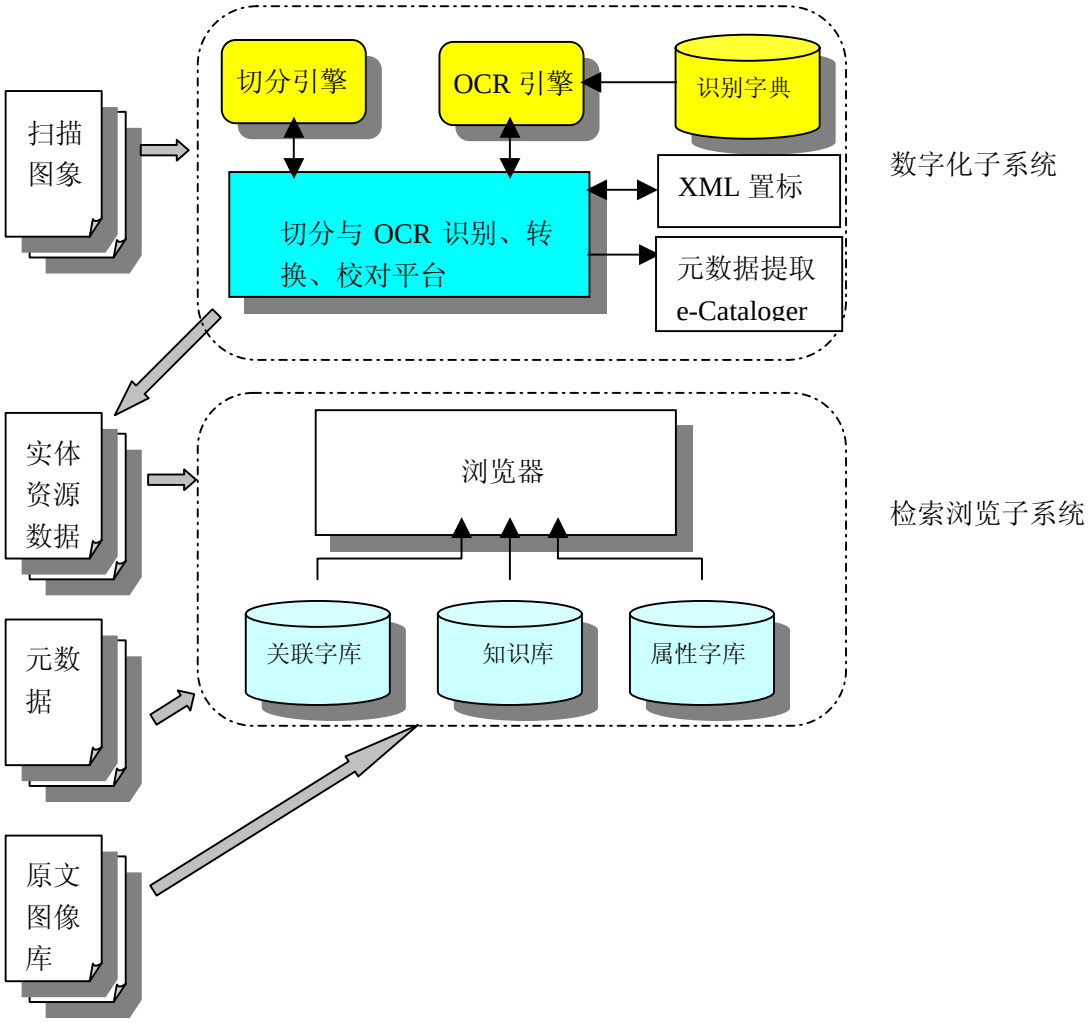
二、 古籍数字化处理流程与体系结构

处理流程：



体系结构：

一个基本的古籍数字化系统至少包括两个子系统：数字化子系统和检索浏览子系统。



三、 采纳先进的技术

工欲善其事，必先利其器。

采用世界上先进、成熟的技术与标准，这是古籍数字化研发成功的保证。

（1）国际标准 ISO/IEC 10646（GB 13000/Unicode）是古籍数字化的适用文字平台

国际标准 ISO/IEC 10646 的全称是：信息技术-通用多八位编码字符集（Information technology-Universal Multiple-Octet Coded Character Set）。在 IT 界另一通俗称呼为 Unicode。

目前已商品化字符集：ISO/IEC 10646 1:2000/Unicode 3.0。

其中包括汉字 27484 个。用户区（EUDC）汉字 5000 余个，共计 32000 余汉字，已成功用于《四库全书》、《四部丛刊》等古籍数字化，但不包括小学类字书用字。

2001 年 11 月正式颁布：ISO/IEC 10646-2:2001（E）/Unicode 3.1，收入汉字七万余个，除甲骨文、篆文外，可满足世界各地汉字使用需要，该标准的商品化也将会实现。

为何古籍文献数字化需要采用 ISO/IEC 10646/ Unicode?

第一、 汉语所表达的丰富内容决定；

第二、文字复杂性：异体字关联的需要。从地域角度包括中、日、韩、越，从汉字本身包括简繁、古今、通假、新旧、正讹等等。

例如：

簡體-繁體關係：简/簡

正體-異體關係：修/修 兔/兎 刃/刃

正字-訛（譌）字：久/乆 派/洏 友/友

通假-被通假：詳/佯

古今字：髡/長

新舊字形：青/青 説/說 媪/媼

中日：卖/売 图/図 单/単

形近異義字：义/叉 刺/刺 諫/諫

避諱字：弘/ 玄/ 燁/ 胤/ 禛/

第三、是古籍国际共享的需要。采用该字符集才可实现一套数据/一套软件，无障碍走向世界。

第四、也是多文种并存的需要。

对于《四库全书》七亿汉字的古籍内容，ISO/IEC 10646 1:2000/Unicode 3.0，可以解决包括小学类（字书、辞书）在内的古籍内容的 99.99%。而且在字型、输入法、编辑器、程序语言、浏览器以及数据库管理系统都获得支持。

《四库》、《四部》用字统计：

《四部丛刊》用字（不含小学类）Chinese Chars Used In SiKuQuanShu

	Hanzi Number	Hanzi Appearence
CJK	18,818	88,990,878
CJK_A	5,183	430,824
CJK_B or Outsider	4,221	1,259,636
Summary	28,222	90,681,338

《四库全书》用字（不含小学类）Chinese Chars Used In SiKuQuanShu

	Hanzi Number	Hanzi Appearence
CJK	18,048	685,901,735
CJK_A	4,900	2,393,467
CJK_B or Outsider	4,212	1,989,121
Summary	27,160	690,284,323

注：有关ISO/IEC 10646 vs. Unicode，请访问http://www.unihan.com.cn中的“实用CJK”栏目http://www.unihan.com.cn/cjk/conception.htm。

（2）采用 OCR 技术，实现图文数码转换

所以选用 OCR 技术实现古籍文字的数码转换，其原因是：对古籍汉字中简繁、异体字的输入，OCR 较之人工录入有优势（十选识别率可以达到 99%，其中的 90%可以正确识别，另外的 9%可以通过点击而不是键盘输入解决）。因此数据加工人员不需再做大量的古籍文字手工录入工作，重点转向文字校对工作。

通过 OCR 可以建立图-文之间形影不离的一一对应关系，便于实现高效率高质量的电脑辅助校对。

有成规模的加工批量。即使初期在 OCR 前后处理的软件研发中要有一定投入，但效率与质量总的效果比人工录入好。

《四库全书》是多特定人按同一格式书写的文字，经过七亿文字 OCR 识别转换、校对的实践，OCR 系统已收到良好实用效果。而经过《四部丛刊》（504 种书）即每种书均为不同书写或印刷版本的复杂多变文字形式的实践和改进，OCR 的适用性和通用性又大为改观。

（3）用软件工具辅助人工校对

怎样在大规模大批量古籍数字化过程中实施高效率、高质量、低成本的文字校对，确保电子出版物的数据质量，这是一个新的课题。

传统的校对方式是校对者在清样与原稿之间反复比较，查错改错。校对过程既要依据纸载体原稿，还要产生大量纸张清样。校对过程是校对者的视觉在清样与底稿之间反复频繁转换，极易疲劳，产生疏漏。而且在初校以后的复校中，校对者还要面对已校的内容（无论正确与否）再次核审，费时费力，若要达到高质量，投入很大。尤其现在，若想聘请一大批懂得古籍文字的人员投入校对工作几乎是不可能的。

我们研发了成套校对软件，辅助一般中等文化水平的年轻人即可完成大规模古籍文字的校对任务，取得了满意效果。

校对作业是在网络环境下在屏幕上进行的。提供的是将古籍原稿的电子图像与数码化的文字对照比较，使校对工作无纸化。其中有页（原稿图像）对页（数码）、列/行（原稿图像）对列/行（数码）、字（取自不同页的原稿图像）对字（取自不同数码页）的形影不离的校对方式，并辅以联机异体字字典，有效地减少了校对者的视觉转移，便利于版面与文字查错，提高工作效率，减少疏漏。同时还提供横向的聚类校对，即把不同页处的同一图像文字取出，看其转换的代码文字是否正确。

为了给总校人员提供有效校对工具，还用数理统计的方法，根据文字识别可信度的统计结果，将易产生差错的字重点提示，将不易产生差错的字隐蔽淡化，使总校工作突出了重点，不仅提高了工作效率，而且使差错率达到低于国家出版行业万分之一的指标。（见图 1—图 12）

（4）采用 XML 作为文献内容的标识语言

XML（Extensible Markup Language）即可扩展标记语言，是一种元语言。它是国际互

联网联盟（W3C）开发的用于网络环境下数据交换、数据管理和网页设计的新技术。它是国际标准 SGML（Standard Generalized Markup Language [ISO 8879]）的一个子集，一个实用标准。

《四库全书》、《四部丛刊》数字化开发的实践表明：XML 非常适合非结构化文献的全文处理，易于表达文献资料；XML 将资料的存贮与显示相分离，可支持同一资料不同格式的显现、输出，支持多种应用程序的处理；XML 可直接应用于因特网，便于开发网络版电子出版物；XML 有良好的层次结构和约束，处理起来很容易，极大地减少软件开发成本；XML 基于资料内容进行标识，因而可被不同程序用于不同用途；XML 具有很强的链接功能，可定义双向链接、多目标链接、扩展链接和文件间链接，非常有利于实现各种关联检索和图文的链接处理；XML 提供了从小配置文件到大规模资料仓库的可扩展性； XML 支持 ISO/IEC 10646/Unicode。

为了实现网络环境下数字图书馆信息共享，对书目、大事记一类元数据，我们还采用 Dublin Core 15 项定义加以规范，并采用 XML 标识，已经收到很好的效果。

（5）实体资源库+知识工具库，实现多种知识信息关联

我们在《四库全书》全文主体数据库的基础上又链接了数字化的《中华古汉语字典》、《四库人名大辞典》、《四库全书简明目录》、“SuperCJK 汉字属性文件”以及“古今纪年换算”等知识库和工具库。这样，《四库全书》电子版不仅可以实现传统的特征检索（题名、作者、关键词词语检索）以及由这些特征构成的布尔组配检索，还可以实现由一个学者到另一个学者、由书目到全文、由著作者条目到其简历、著述、由相关作者到相关作品的多个知识点关联检索，还可实现对读者不解的文字立即提供释义与读音，实现文中古代纪年表示对公元纪年的换算等。

所有的努力都是朝着一个目标：最大限度地发挥电子出版物的优势、最大限度地满足读者的需求。把知识联系起来，把书本上的内容贯穿上电子的经络，实现全文检索、字字可查，句句可检，帮助读者进行快速的非顺序式的检索、实现由此及彼、由表及里的查阅。

（6）中日、简繁、异体汉字关联

从事古籍数字化开发，不仅着眼于中国大陆，也着眼于台、港、澳、日地区以及世界各地华人和研究中国文化的读者的需求。因此简、繁、异、日等各种汉字关联转换功能，可使仅熟悉某种类型汉字的读者（如只熟悉简化字或只熟悉繁体字）在检索时能“简入繁出”、“繁入简出”、“正入异出”、“异入正出”、“日（日本汉字）入中出”……均可方便地查到所需文献。

（7）词语有限范围控制查询帮助学者准确捕捉所需资料

全文检索优点突出，但并不完美。尤其是读者为了快速准确获取所需主题资料在进行词语组配检索时，若不加以控制，命中结果大部分可能不是自己所需，这点在古籍查询时尤为突出。

为此，我们提供了一下有限范围控制手段：

- 词语间限定一定字数
- 词语限定在一定的类别
- 词语限定在一个段落内
- 词语限定在指定作者的作品中
- 词语限定在特点作品（书名）中

事实证明采取这些限制命中率大大提高。

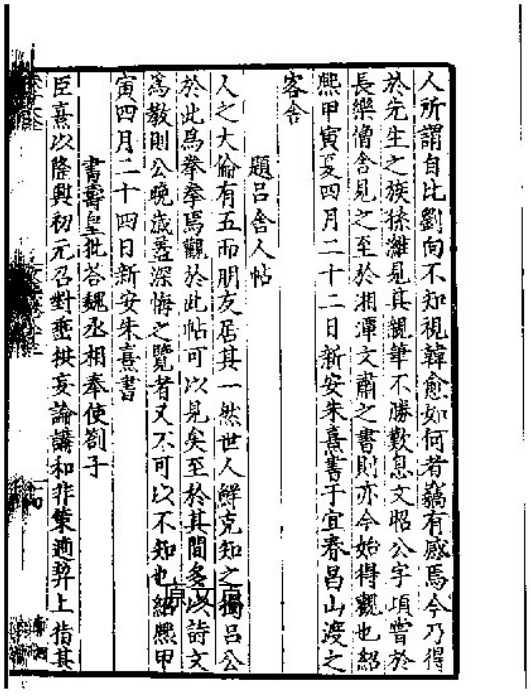
（8）全球版、网络版应是古籍数字化产品的重点，还应纳入数字图书馆系统

对于具有局域网络的大学、研究团体用户，应提供古籍局域网络版产品；对于只想通过 Internet 利用古籍的用户，应提供 Internet 版使用方式。这样做也利于版权保护。当前还应注意的是，要使古籍、包括大部头丛书中的一个个单册，纳入数字图书馆系统，这已经提到日程。

参考文献：

1. 张轴材 朱岩：《数字化内容与数字化工具》（2000 年 10 月）
2. 朱岩：谈古籍数字化（《两岸三地古籍与地方文献》2002 年 2 月）
3. 王晓波：《大规模古籍电子化关键技术及实现》（2000 年 6 月）

图 1


人所謂自比劉向不知視韓愈如何者藉有感焉今乃得於先生之孫猻潛見其親筆不勝歎息文昭公宇頃嘗於長樂僧舍見之至於湘潭文蕭之書則亦今始得觀也紹熙甲寅夏四月二十二日新安朱熹書于宜春昌山渡之客舍

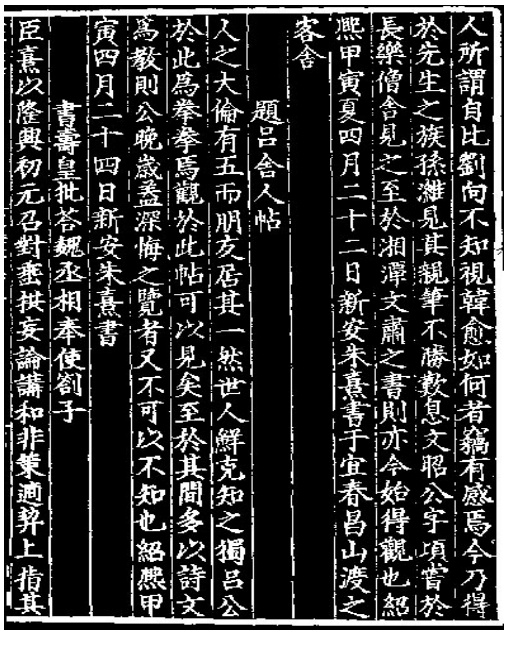
題呂舍人帖

人之大倫有五而朋友居其一然世人鮮克知之獨呂公於此爲拳拳焉觀於此帖可以見矣至於其間多以詩文爲教則公晚歲益深悔之覽者又不可以不知也紹熙甲寅四月二十四日新安朱熹書

書壽皇批答魏丞相奉使劄子

臣熹以窶與初元召對亟拱妄論講和非策適幹上指其

图 2


人所謂自比劉向不知視韓愈如何者藉有感焉今乃得於先生之孫猻潛見其親筆不勝歎息文昭公宇頃嘗於長樂僧舍見之至於湘潭文蕭之書則亦今始得觀也紹熙甲寅夏四月二十二日新安朱熹書于宜春昌山渡之客舍

題呂舍人帖

人之大倫有五而朋友居其一然世人鮮克知之獨呂公於此爲拳拳焉觀於此帖可以見矣至於其間多以詩文爲教則公晚歲益深悔之覽者又不可以不知也紹熙甲寅四月二十四日新安朱熹書

書壽皇批答魏丞相奉使劄子

臣熹以窶與初元召對亟拱妄論講和非策適幹上指其

去污、端正

图 3

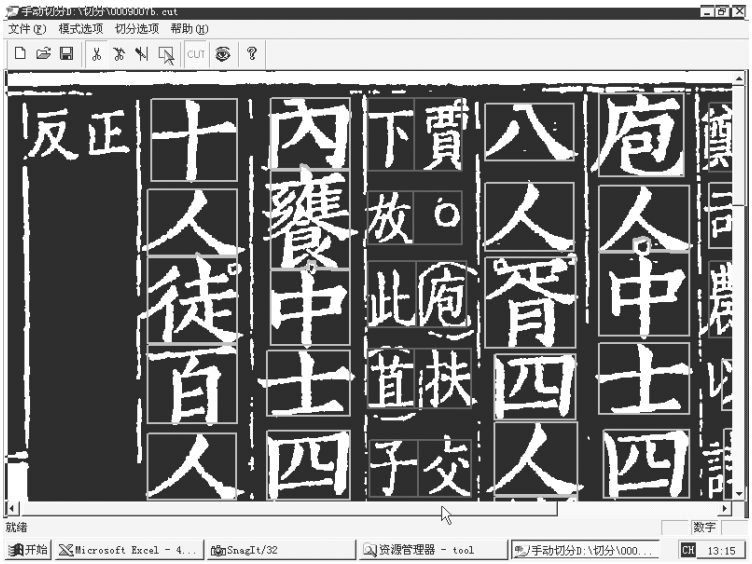
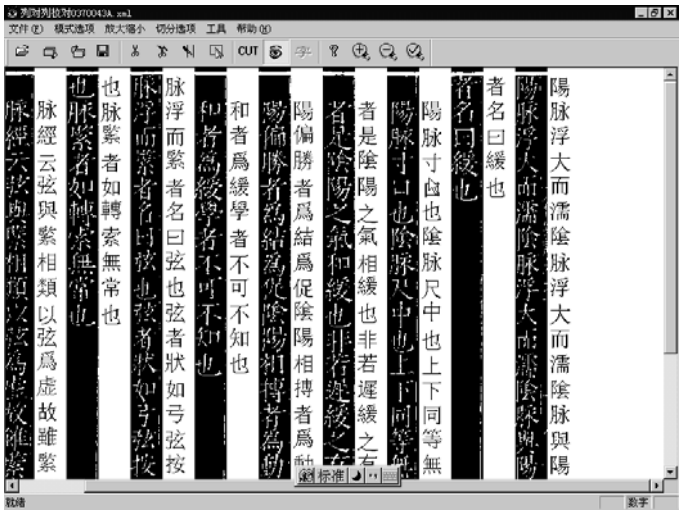
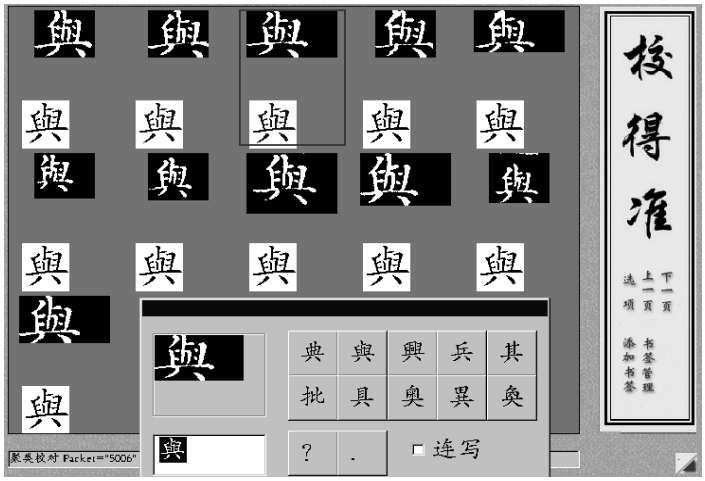


图 4



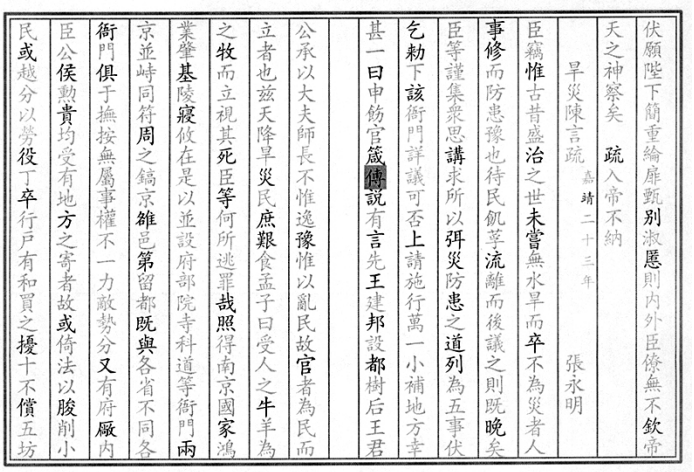
列对列校对

图 5



聚类校对

图 6



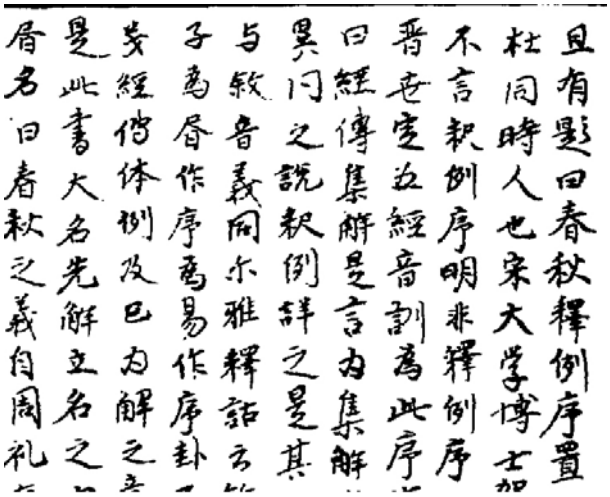
伏願陛下簡重綸扉甄別淑慝則内外臣僚無不欽帝天之神察矣　疏入帝不納

旱災陳言疏　嘉靖二十三年　張永明

臣竊惟古昔盛治之世未嘗無水旱而卒不為災者人事修而防患豫也待民飢莩流離而後議之則既晚矣臣等謹集衆思講求所以弭災防患之道列為五事伏乞勅下該衙門詳議可否上請施行萬一小補地方幸甚一曰申飭官箴儆諛說有言先王建邦設都樹后王君公承以大夫師長不惟逸豫惟以亂民故官者為民而立者也茲天降旱災民庶艱食孟子曰受人之牛羊為之牧而立視其死臣等何所逃罪哉照得南京國家鴻業肇基陵寢攸在是以並設府部院寺科道等衙門兩京並峙同符周之鎬京雒邑第留都既與各省不同各衙門俱于撫按無屬事權不一力敵勢分又有府厥内臣公侯勳貴均受有地方之寄者故或倚法以朘削小民或越分以勞役丁卒行戶有和買之擾十不償五坊

重点提示校对

图 7



且有題曰春秋釋例序置杜同時人也宋大学博士加不言釈例序明非釋例序晉吉定五經音訓為此序曰經傳集解是言為集異同之說釋例詳之是其与敘音義同尓雅釋詁子為書作序為易作序卦爻經倣体例及已為解名曰春秋之義自周礼

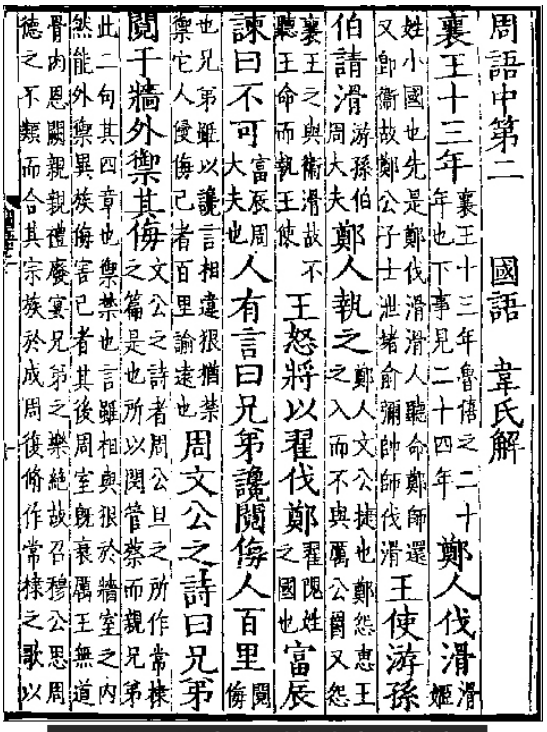
《春秋正义》日本正宗寺钞本原书页

图 8

且有題曰春秋釋例序杜同時人也宋大学博不言釈例序明非釋例晉吉定五經音訓爲此曰經傳集解是言為集異同之說釋例詳之是与敘音義同尓雅釋詁子爲書作序爲易作序義經傳体例及已为解

《春秋正义》日本正宗寺钞本转换后文本页

图 9



周語中第二　國語　韋氏解

襄王十三年襄王十三年魯僖之二十年也下事見二十四年鄭人伐滑滑姬姓小國也先是鄭伐滑滑人聽命鄭師還又即衛故鄭公子士泄堵俞彌帥師伐滑王使游孫伯請滑游孫伯周大夫鄭人執之鄭人文公捷也鄭怨惠王之入而不與厲公爵又怨襄王之與衛滑故不聽王命而執王使　王怒將以翟伐鄭翟隗姓之國也富辰諫曰不可富辰周大夫也人有言曰兄弟讒鬩侮人百里鬩侮也兄弟雖以讒言相違狠猶禁禦它人侵侮己者百里諭遠也周文公之詩曰兄弟鬩于牆外禦其侮文公之篇是也所以閔管蔡而親兄弟此二句其四章也禦禁也言雖相與狠於牆室之內然能外禦異族侮害己者其後周室既衰厲王無道骨肉恩闕親親禮廢宴兄弟之樂絕故召穆公思周德之不類而合其宗族於成周復脩作常棣之歌以

《国语》原书页<杭州叶氏藏本>

图 10

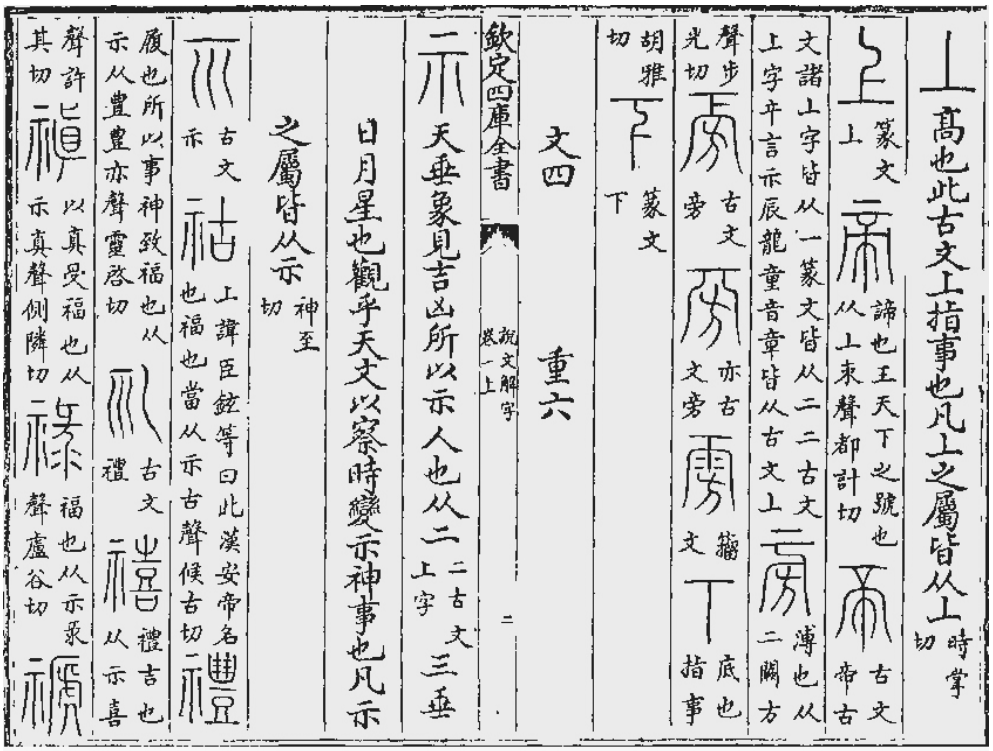
周語中第二　國語　韋氏解

襄王十三年襄王十三年魯僖之二十年也下事見二十四年鄭人伐滑滑姬姓小國也先是鄭伐滑滑人聽命鄭師還又即衛故鄭公子士泄堵俞彌帥師伐滑王使游孫伯請滑游孫伯周大夫鄭人執之鄭人文公捷也鄭怨惠王之入而不與厲公爵又怨襄王之與衛滑故不聽王命而執王使　王怒將以翟伐鄭翟隗姓之國也富辰諫曰不可富辰周大夫也人有言曰兄弟讒鬩侮人百里鬩侮也兄弟雖以讒言相違狠猶禁禦它人侵侮己者百里諭遠也周文公之詩曰兄弟鬩于牆外禦其侮文公之篇是也所以閔管蔡而親兄弟此二句其四章也禦禁也言雖相與狠於牆室之內然能外禦異族侮害己者其後周室既衰厲王無道骨肉恩闕親親禮廢宴兄弟之樂絕故召穆公思周德之不類而合其宗族於成周復脩作常棣之歌以

四部叢刊　國語

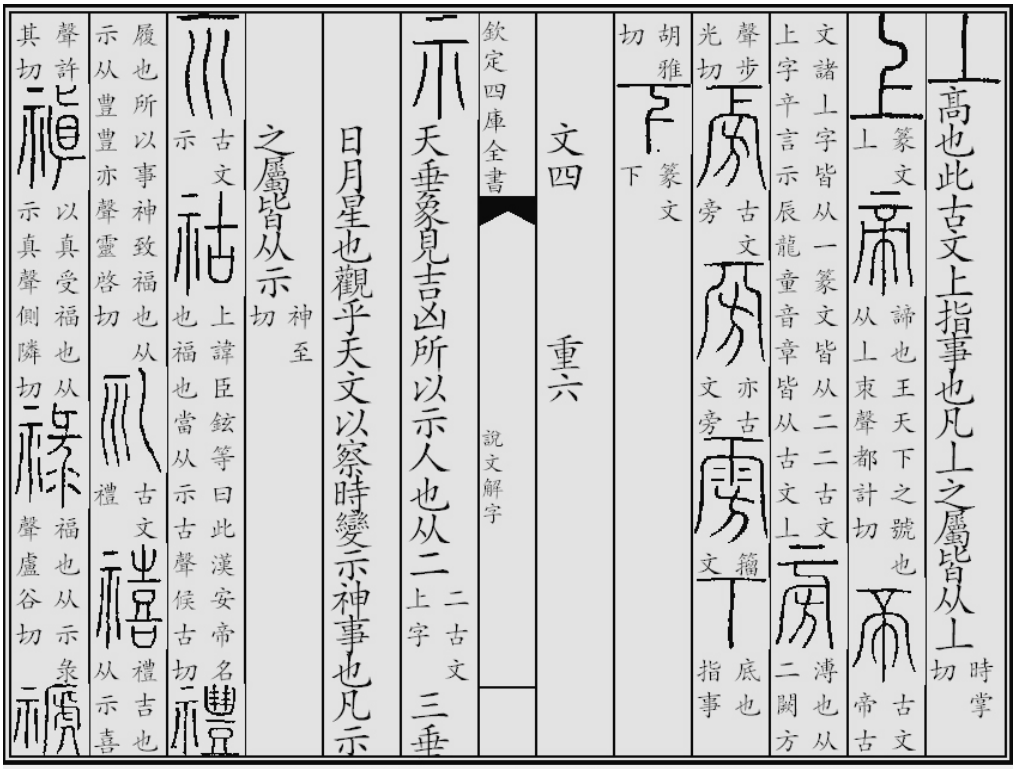
《国语》转换后文字页

图 11



《说文解字》（四库全书）原书页

图 12



《说文解字》（四库全书）转换后文本页