由中国人民大学信息资源管理学院冯惠玲教授、加小双副教授主讲的《数字人文导论》是中国人民大学本科生“数字人文荣誉辅修学位”系列课程之一，选课学生涵盖本、硕、博三类。课程内容包括数字人文导论（理论讲解、案例分享）和数字人文项目设计（实践操作）。

本系列推文共12期，是该课程学生课堂展示的成果报道。第四期由信息资源管理学院2019级本科生李慧和国学院2019级本科生王鑫共同完成，以台湾大学资讯工程学系项洁教授主持的“DocuSky数位人文学术研究平台”项目为研究对象，对该项目的背景、内容设计、实现路径、项目成果和特色进行系统梳理，最后对该项目进行评析，总结对今后数字人文项目的启示。

Docusky数位人文学术研究平台

概况

平台工具简介

图1 工具界面

l 标记与编辑：为需要研究的语词添加标记(tag)，以便日后浏览和统计分析；

l 转换文本格式：将表格、 纯文字、MARKUS等格式的文本与诠释资料转换为DocuXML格式，以便上传至云端的个人资料库和后续的统计、分析、可视化；

l 建库与重整：将本地DocuXML文件上传至云端，并进行整理；

l 探勘与分析：对资料进行各式分析和统计，如统计词频、分析语言风格、撷取同类型词汇等；

l GIS与视觉化：允许用户查询历史地名、整合地理信息、绘制多层地图。

平台资源简介

图2 资源关联

平台上的资源主要是外部资源，包括：

l CBETA:汉文佛典
l CTEXT:中国哲学书电子化计划
l Kanripo:汉籍资料库
l WikiSource:发布处于公有领域或者法律上允许自由发布的原文本
l CBDB:中国历代人物传记资料库

平台上的工具和资源目前仍在不断完善中。

平台架构

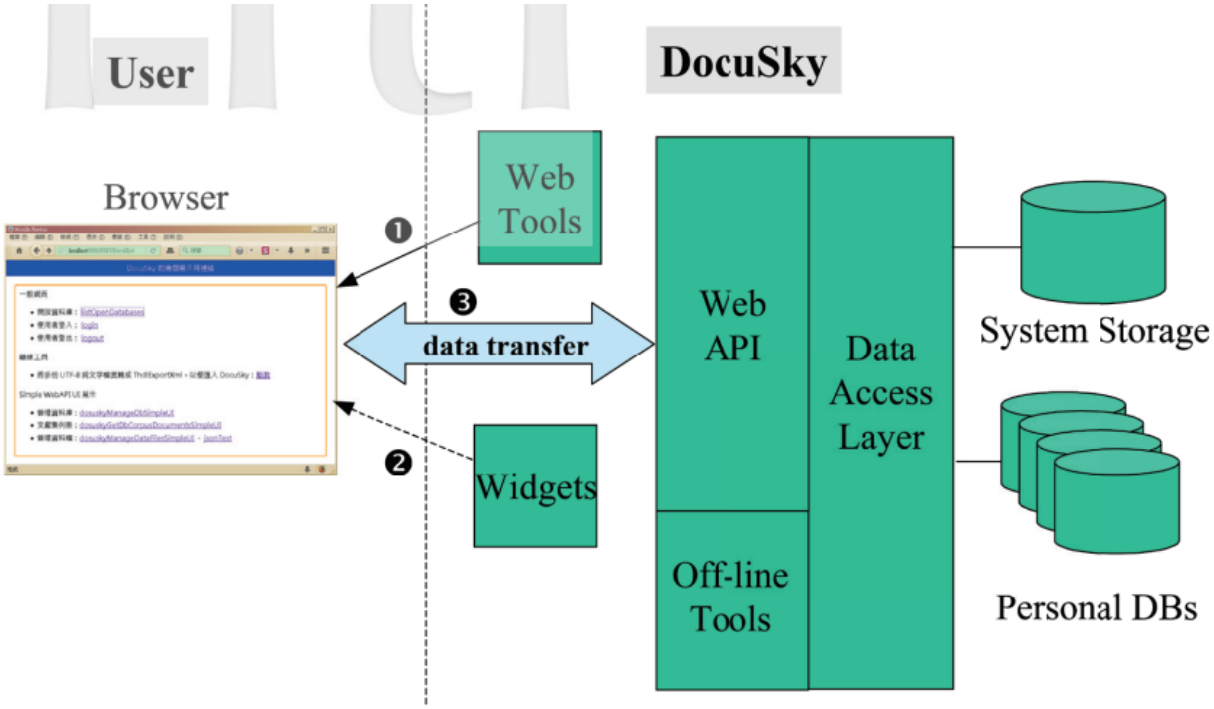

图3 平台结构示意

左侧是客户端，使用者通过浏览器进行操作；右侧是服务端。其中，web tools是上文提到的五类工具；使用者通过工具操作个人资料库（Personal DBs），这个操作过程由中间层（Web API，Off-line-Tools,Data Access Layer）实现。

特色

1.通用性。设计开发者杜协昌教授指出：“工具开发时，不必然假设需运用在某份特定的文本上”，也就是说，工具对所有文本都适用，工具具有通用性。

2.个人化。正如平台“研究目的”所说，该平台让人文研究者不必再步步仰赖信息科技专家，自主且自由地融合数字科技进行人文研究。

3.标准化。杜协昌教授：“因为使用者的文本涵盖了文字内容、诠释资料以及标记信息，而每位研究者关心的主题和研究取径又各自不同，我们从方法论的角度出发，认为必须设计一套通用且有弹性的格式（DocuXml），才能对多样化的文本内容进行妥善的处理”。DocuSky平台在XML（可扩展标记语言）的基础上，设计了DocuXML格式，能够对文字内容、诠释资料和标记信息进行妥善的处理。

评析

适用方向

宫词不是主流的诗歌创作，历代诗话鲜有论及，这就导致在阅读宫词时难以把握不同作家与作品间艺术水平的高低。又由于题材的单调重复，人工阅读体验效率低下，适于交给机器处理。

陈汉文在研究宋白《宫词》百首时即应用了DocuSky进行词频统计和文风分析。图3及表1来自其论文《论宋白〈宫词〉百首的字词运用及其文学意义》，讨论宋白《宫词》百首对宋人清雅诗风的影响。

(a) 北宋・宋白〈宮詞一百首〉　(b) 唐代・王建〈宮詞一百首〉

(c) 北宋・王珪〈宮詞一百首〉

图4 宋白、王建、王珪常用词对比

表1 宋白及王建的单音节词使用频率对比

宋白宮詞 1 字詞	頻率	百分比（%）	王建宮詞 1 字詞	頻率	百分比（%）
宮↑	56	100.00	金↑	42	100.00
春↑	41	73.21	中↑	40	95.24
花	29	51.79	人↑	39	92.86
金↓	28	50.00	來↑	37	88.10
玉	28	50.00	一↑	33	78.57
新	23	41.07	頭↑	30	71.43
一↓	22	39.29	殿↑	30	74.46
天	21	37.50	花	29	69.05
中↓	21	37.50	紅↑	27	64.29
不↓	20	35.71	不↑	27	64.29

資料來源：作者自行整理。

註：以「宋白宮詞 1 字詞」為例，紅色指其在宋白〈宮詞〉使用率較王建高；綠色指其在宋白〈宮詞〉使用率較王建低。至於「王建宮詞 1 字詞」，紅色部分說明此字的使用頻率較宋白高。

又如丁声树《诗经“式”字考》。丁声树考证的大致思路是“对于一个未确定意义的词，如果它出现时，另一个词也连带着经常出现，那么二者之间很可能存在意义上的联系”，从描述上便可知适用词频统计。在《诗经》文本中提取出所有含“式”的句子，再统计这些文本的词频，再与原始文本词频进行比较（见表2）。快速发现“式”与“无”的对应关系，确定训诂。

表2 《诗经》词频对比

文獻集高頻詞 (grams) 比較 ▲

全文语料.txt:1gram

文字雲 Export (csv)

No.	Gram	DocFreq	TermFreq
1	之	1	2695
2	不	1	1714
3	有	1	1600
4	其	1	1474
5	我	1	1379
6	於	1	1097
7	子	1	1091
8	以	1	888
9	維	1	880
10	無	1	846

文獻集高頻詞 (grams) 比較 ▲

诗经式语料.txt:1gram

文字雲 Export (csv)

No.	Gram	DocFreq	TermFreq
1	式	1	50
2	無	1	14
3	不	1	14
4	之	1	10
5	以	1	8
6	燕	1	7
7	爾	1	7
8	酒	1	7
9	君	1	7
10	俾	1	7

批判分析

该工具（见图5）没有识别词性的功能，导致无法满足句法方面的要求。其次，面对“釐”、“赍”等词时可能因为义项较为生僻，工具识别不到所以忽略，但是这些例句在研究过程中的价值是并不因为生僻而降低。

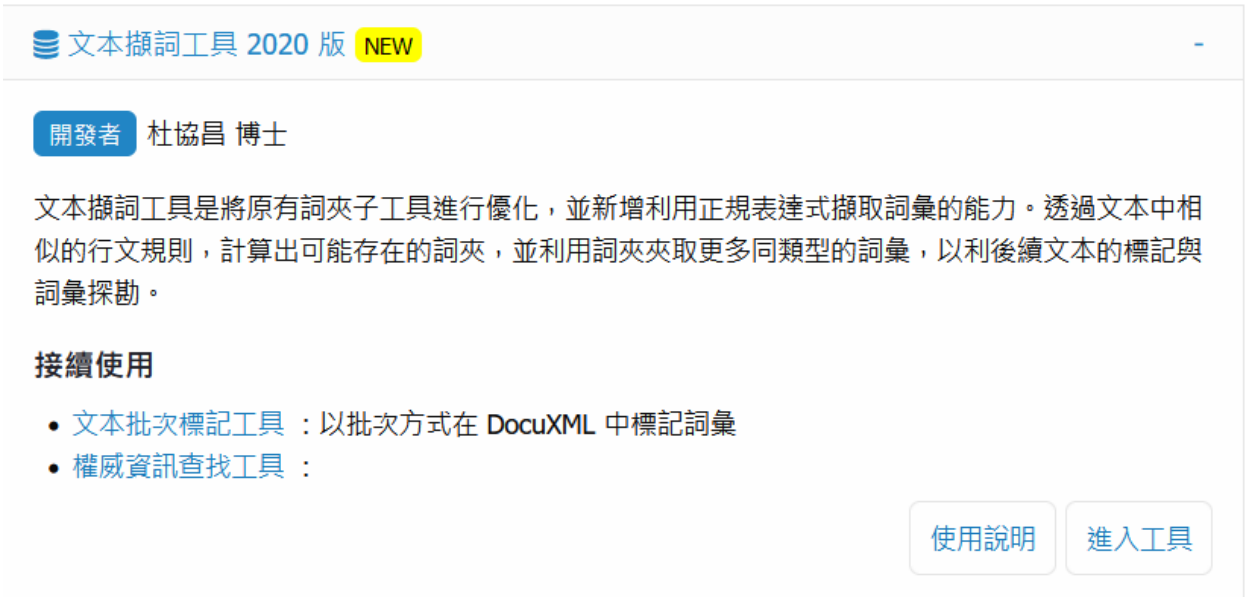

图5 文本擷词工具

除了历史研究，经常用到对读方法的是语言学。美中不足的是，docusky所提供对读系统（见图6）目前仅支持相同语言的对比。梵汉、藏汉对音等多语种比较尚属阙如。
但docusky所关联的、由法鼓文理学院开发的dedu工具则支持梵汉藏巴多语言对读。

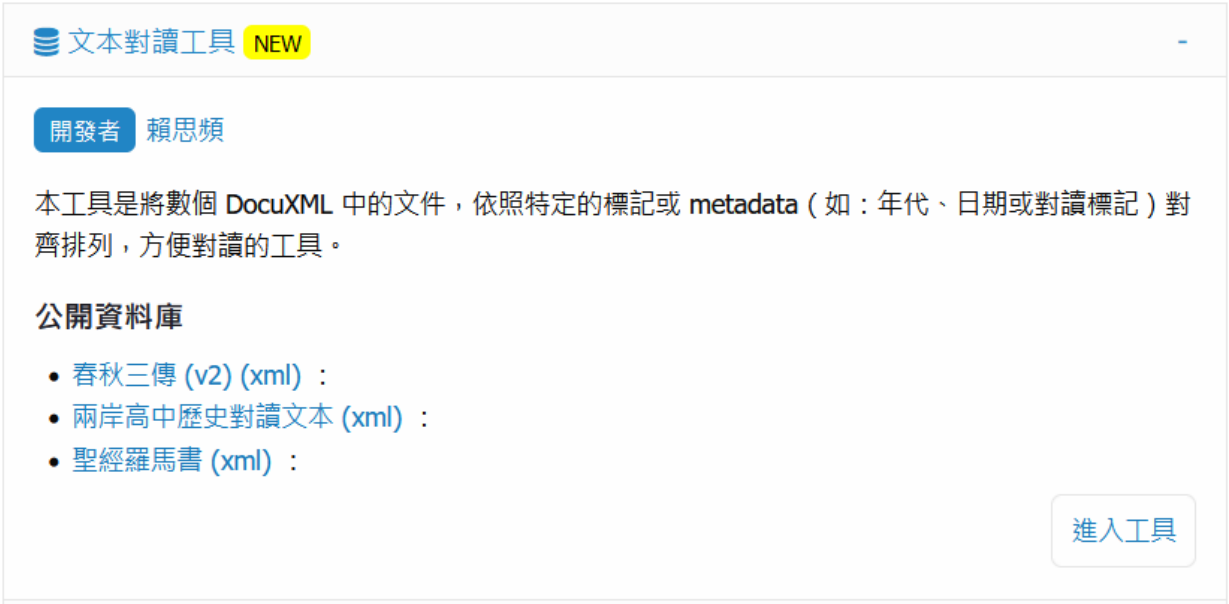

图6 文本对读工具

总结

利用此类工具回望过去的学术成果，会发现很多研究方法都可以交给机器执行，从而集中精力于一途。该平台试图让人文学者自由自主地利用什伯之器而不必仰求于人。
物无全美，该平台所提供的诸工具还有不少值得改进之处。例如本平台所提倡的通用性是基于文学、史学的通用性，于语言学注目甚少。同时，标注的层次和深度有待拓展，目前其标记库仅支持人名、地名和职官等。数据分析的数量存在限制，词频统计功能语料超过十万字仅能输出乱码。
瑕不掩瑜，DocuSky涉及数字人文的诸多方面，无论是优点还是缺陷，都给日后的项目带来宝贵的经验。

参考文献

[1]杜协昌.DocuSky：个人文字数据库的建构与分析平台[J].数位典藏与数位人文,2018(2):71-90.
[2]陈汉文.论宋白《宫词》百首的字词运用及其文学意义[J].数位典藏与数位人文,2021(7):1-36.
[3]丁声树.诗经“式”字说[J].历史语言研究所集刊第六册,1936:487-495.

编辑：李慧 王鑫
排版：李慧 王鑫

公众号账号：rucdh2019

网址：http：//dh.ruc.edu.cn

邮箱：rucdh@ruc.edu.cn

中心简介

中国人民大学数字人文研究中心集人民大学多学科优势，秉持融合文理、协同创新之理念，开展数字人文理论研究、实践探索、人才培养和学术交流。