刘炜
Wei Liu

刘炜，上海图书馆上海科学技术情报研究所副馆（所）长，研究员，图书情报学硕士，计算机软件与理论博士。复旦大学、华东师范大学和上海大学兼职教授，上海大学博士生导师。兼任上海市图书馆学会副理事长，DC 元数据管理委员会委员，《中国图书馆学报》等专业期刊编委。

Wei Liu is the Deputy Director of Shanghai Library. He is an adjunct professor of Shanghai University, as a doctoral tutor and give lectures on Digital Libraries and Digital Humanities. He also serves as the inaugural director of the China Digital Humanities Alliance (under the Chinese Society of Indexing).

王丽华
Lihua Wang

王丽华，上海大学图书情报档案系副主任，副教授，图书馆学博士，上海大学数字人文研究与发展中心主任，《数字人文研究》编委。

Lihua Wang, Doctor of Library Science, is deputy dean and associate professor at Department of Library, Information and Archives of Shanghai University, and the director of Digital Humanities Research and Development Center. She is also the editorial board member of Digital Humanities Research.

汉学数字人文通用平台的需求与设计

The Functional Requirement and Designing of a Universal Digital Humanities Platform for Sinology

刘炜
Wei Liu
上海图书馆副馆长
Deputy Director, Shanghai Library

王丽华
Lihua Wang
上海大学图书情报档案系
Department of Library, Information and Archives, Shanghai University

摘要

随着人文社会科学向数据驱动型研究范式的转型，图书信息机构提供数字人文服务已成趋势，很多数字典藏系统都在积极应用关联数据、知识图谱、实体识别、机器学习、数据可视化等数据技术，升级为数字人文服务平台。文章讨论了汉学数字人文平台建设的功能需求和设计趋势，从系统、资源、功能、工具与用户五个方面探讨了分布式环境下全球汉学数据典藏机构进行平台共建共享的系统要求、资源规范、功能特征、工具支持与用户界面等整体需求，对数字人文平台为支撑数字人文研究而提供全面的资源获取能力、个人研究环境和数据操控工具等，提出了较为系统完整的设计思考，并以上海图书馆正在开发中的“历史人文大数据平台”为案例阐述这些思考成果的应用。

关键词：汉学研究、数位人文、数据资源、平台设计、需求分析

Abstract

With the transformation of the humanities and social sciences into a data-driven research paradigm, it has become a trend for libraries and other memory institutions to provide digital humanities services. Many practitioners are actively applying data technologies such as Linked Data, Knowledge Graph, Name Entity Recognition, Machine Learning, and Data Visualization

to upgrade digital collection systems to Digital Humanities Service Platforms. This paper discusses the functional requirements and design thinking of a Digital Humanities Platform for Chinese Studies. It also discusses the data and metadata formats, protocols for data exchange, and possible mechanisms for co-construction and sharing of platforms by global sinology repositories in a distributed environment from the aspects of system, resources, function, methods and users. At the same time, the paper proposes a systematic thoughts on how to adapt the digital humanities platform to meet the needs of digital scholarship and provide comprehensive resource access, personal research environment, and data manipulation tools to researchers. The results of these reflections are being applied to the construction of the "History and Humanities Data Platform" currently developed by Shanghai Library.

Keywords: Sinology Study, Digital Humanities, Data Resources, Platform Design, Requirements Analysis

預錄及簡報檔
https://pse.is/3e37pp

壹、引言

人文学科是所有科学之肇始，是人文精神之依托，是被称为知识分子的必备和基础素养。无论是古希腊的七艺（文法、修辞、逻辑、算数、几何、天文、音乐），还是春秋的六艺（诗、书、礼、乐、易、春秋），其所创立的知识教育体系在今天多数归属于人文学科的范畴，致力于培养区别于万物的所谓“人性”。而当今社会建立起与工业文明相匹配的极其复杂又高深的现代教育，看似造就了大量知识丰富的“专家”，但却带来了知识分子整体上的灭绝，不仅缺乏对人的价值以及人类未来命运的思考者，连培养基本的责任与担当都成了奢望。在这个机器智能和生命编辑的时代，人文主义越来越遭遇危机，乃至面临灭顶之灾，我们比任何时候都更加需要和呼唤世界意义的守护者 [1.]

正是这样一个背景，诞生了数字人文。

作为信息技术在人文领域的应用，数字人文目前仍处于非常早期的发展阶段。虽然其历史可以追溯到计算器刚开始用来做文字处理的上世纪中叶，迄今已有七十余年，但“数位人文”一词是 2004 年随着《A Companion to Digital Humanities》一书的出版才得以定名，当前还不具有公认的定义，甚至连边界在哪里也众说纷纭、莫衷一是。即便如此，鉴于数字化社会已势不可挡，印刷品不再是知识生产与传播的主要媒体，Tony Hey 等提出“科学研究第四范式”即“数据驱动型研究”概念 [2.]，当所有的研究素材和方法都数字化之后，人文科学也概莫能外，数字人文必然是人文研究的未来。

数字人文是各门具体人文科学采用数字方法的汇聚和总结，是一种“方法论共同体（Methodological Commons）” [3.]。然而目前这个共同体还不具备库恩所说的共同的“学科范式”特征，还是各种方法、技术的大杂烩，没有形成一定的结构和规律性认识，该领域的研究者甚至对数位人文能不能成为一门“学科”还心存疑虑而争论不休。当然近二十年来数位人文的研究基础设施已得到了很大程度的建设和完善 [4.]，拥有了颇具影响力的协会、学会和专业期刊，定期召开国际或地区性会议，具有稳定的基金支持，尤其形成了本 - 硕 - 博的专业教育体系，目前的薄弱环节是数据资源相关的平台建设和标准规范尚未成熟，方法论体系也没有发展成型并得到公认。

1. 尤西林 . 闡釋並守護世界意義的人 -- 人文知識份子的起源及其使命 [M]. 上海：華東師範大學出版社，2017：1.
2. Tony Hey 等著，潘教峰等译 . 第四範式：數據密集型科學發現 [M]. 北京：科学出版社，2012：i.
3. McCarty W. Humanity Computing. [EB/OL]. http://www.mccarty.org.uk/essays/McCarty,%20Humanities%20computing.pdf
4. 劉煒等 . 面向人文研究的國家數據基礎設施建設 [J]. 中國圖書館學報 2016.5.29-39.

以汉学为代表的中文数位人文研究也处在一个刚刚起步的阶段。早期的数位图书馆或数位典藏成果为当下的数位人文研究提供了重要的数据支持，然而从整体上看仍不系统，缺乏规划，各学科发展也很不平衡，研究成果较为零散、微观，多是对数位技术的简单应用、对过去研究的重复验证，或者是对西方研究的一种单纯模仿，还缺乏有影响力的、独创性的成果。究其原因，图书馆等人类记忆机构在数据基础设施建设方面的滞后是一个重要瓶颈[5.]。相比西方国家，我们在数据获取方面的困难要大得多：数据系统之间缺乏联通，付费墙壁垒高耸，造成数据获取的不充分和不完整，或者缺乏必须的数据格式（如中文文献大多以图像方式提供，文本奇缺），影响到项目的成本、成果的水平，以及对数字人文研究方法的归纳总结和教育机构相关人才的培养等，已成为汉学数字人文发展的严重制肘。

本文试图探讨一种开放的数字人文服务平台的设计，不仅满足一般人类记忆机构将数字典藏系统升级为基于数据的服务设施，重点在灵活可迁移的云平台架构设计，以及可互操作、热插入，容器化的应用 app 生态建设，同时考虑所有机构平台之间的互联互通协议和方案，以及如何应用关联数据、知识图谱、实体识别、机器学习等技术，提供人文研究各类文本、图像、社交网络、地理信息和可视化等通用工具的支持，长远支持数字人文项目的全生命周期管理。

貳、数字人文研究与数字人文平台建设

一、 传统人文研究与数字人文研究

人文研究一般是人文学者针对特定问题，综合利用各种材料，透过一定方法，经过研究过程而得出结论并发表交流的完整流程。传统人文研究的素材可分为实物、文献（文本或图像）和抽象物（概念、角色等）等，依靠思考和写作完成研究，整个研究过程融合了多种研究方法，是一个从资料收集，到分析归纳，最后得出结论的线型过程。

数字人文研究也基本如此，但稍有不同的是，其“原料”对象可以进一步区分为数字文本、数码图像或由数字对象构成的一个“模型”，有学者称之为“数据态”，这个模型可以很复杂，复杂到作为某个真实系统的仿真（即所谓数字双生 Digital Twins）；其研究方法可以包括传统方法的计算器实现，或任何新颖的计算方法（统计、分析、聚类、可视化）。数字人文研究方法可以具体分类为技术、过程和行为三个方面，其过程也比传统人文研究的线型过程要复杂得多，可以是来回反复的交互过程，其成果发表和交流形式也多利用网络或社交

5. 包弼德等 . 数字人文与中国研究的网络基础设施建设 [J]. 圖書館雜誌 2018.11.18-25.

媒体，具有迅速、便捷、容易追踪但转瞬即逝的特点，且目前还没有[6.]很好的计量与评价方法。

不同的人文学科在迈向数字人文过程中其资源特征和方法特征都有所不同，例如文学或语言学偏重于利用文本处理技术，历史学则关注实体对象的时空呈现及相互关系，哲学需要将文本抽象为特定语义的概念，等等，可以看成是数字人文上述细分要素不同配方的“跨模态”组合。当然这个讨论还只是数字人文研究一般方法的一个思考框架（如图 1 所示），目前无论是具体的人文学科，还是一般性的数字人文，其方法体系都没有定型，还处在发展变化中，也有待进一步挖掘整理。

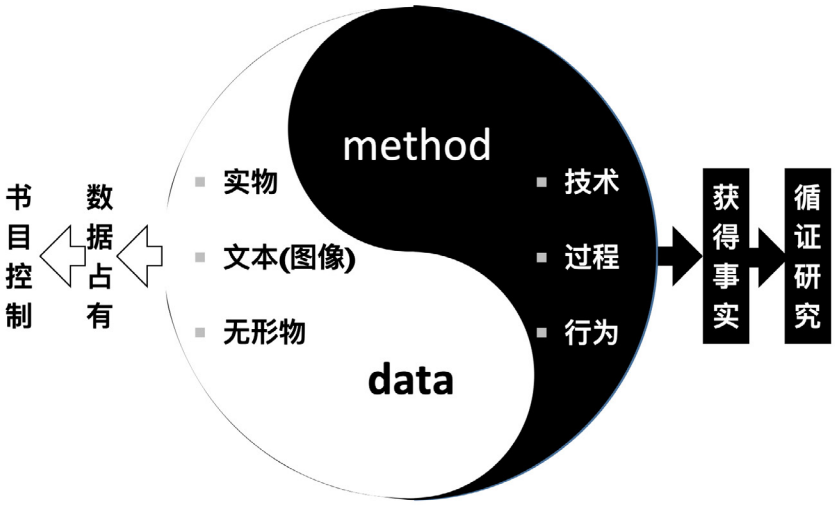

图 1 数字人文研究的要素组成：数据与方法

二、两者的不同

(一) 在研究过程方面

传统人文研究对于素材的收集、加工、处理是研究过程的开始，这是人文研究很重要的有机组成部分，而数字人文可以将数据汇集、处理的通用部分独立出来，作为研究基础设施的一部分，由专门的图书馆、档案馆等相关机构去完成，这就区分了基础设施建设工作和数字人文研究工作。目前数字人文领域大量的工作其实是基础设施建设工作，可以看到中文期刊数字人文的论文发表中大量来自图书馆信息档案学科，就是这个道理。但基础设施建设并不能代替数字人文研究，前者的目的是为了促进后者。

6. 王曉光“新技术”和“文科”不能简单相加 光明日报 2020.12.29.p14. https://epaper.gmw.cn/gmrb/html/2020-12/29/nw.D110000gmrb_20201229_3-14.htm.

(二) 在素材内容方面

传统人文通常透过管理和操控载体化的文献取得内容，限于手工处理的效率，研究的广度、深度都受到限制；而数字人文研究基于数据，平台通常就提供细粒度的知识组织，甚至建立了语义联系，使得材料的操控变得较为容易，能够进行更大范围深入研究，跨学科研究也更为容易。

(三) 在研究方法方面

传统人文研究大都采用定性的思辨方法，透过联想、比较、逻辑推理、思想实验等进行叙事或阐释；而数字人文可以采用建立模型和定量方法，进行文本分析、内容分析、时空分析、社会关系分析、统计聚类、可视化展示等，从某种程度上为人文研究提供了一定的可重复可验证的科学性保证。

(四) 在采用的技术方面

传统人文研究可能会采用田野调查、问卷访谈等；而数字人文可以透过各种技术，如机器学习、神经网络、语意标注、文本分析、量化分析、聚类算法等等。

(五) 在科研协作方面

传统的人文研究大多是学者个人或小规模团队透过多年皓首穷经、苦思冥想，坐多年冷板凳忽然顿悟，取得些许进展；而数字人文更强调大规模协同和社会网络交互，甚至大量采用众包方式，网络平台能否提供相应能力就显得非常重要。

(六) 在成果交流方面

传统人文基本上以出版图书或发表论文为最高标准；而数字人文可以同时推出网站、数据集、工具、软件、课件、博客文章、可视化作品、多媒体电子书等，专著和论文可以只是的副产品。当然数字人文对于基础设施可以更丰富和全面，包含计算设施、云平台、资源库、语料库等等。

三、数位人文平台

数字人文平台是为数字人文研究服务的，也是数字人文研究最重要的基础设施之一，平台建的好不好最终是要透过数字人文研究成果来检验的。因此在建立之初需要了解数字人文研究人员的需求，了解数字人文研究的一般规律，以及方法、过程和行为，否则也无法设计

出好的数字人文平台。当然，数字人文平台首先需要“兼容”传统的人文研究，这种需求在很大程度上数字典藏系统（数字图书馆）就能满足，然后可以进一步升级开发“真正的”数字人文平台，向人文学者全面提供基于数据的研究基础设施。

目前的数字典藏系统是一种初级版本的数字人文平台，由于其大都只是将传统的文献扫描成图像，结合元数据库提供有限途径的查询，功能十分有限，基本上只是传统图书馆的一种载体转换而已，无法满足数字人文研究的进一步需要。虽然有一些平台已开始提供一些工具，例如分词、标点、批注、词云、格式转换、实体提取、人物关系呈现及可视化等，并采用了众包理念，但总体上还较为简单，集成了一些成熟度不一的功能，没有结合人文学者的领域和场景，用户体验不够好。

现有的数字人文平台存在的最大问题还是技术上的，在内容管理上尚未采用知识图谱为代表的语意数据管理技术，还是关系数据库或者全文数据库；在体系结构上虽然已注意借鉴云计算技术，但还没有充分考虑以微服务和容积技术为基础的云原生架构，也没有考虑技术架构和内容架构分离的设计。因此很难满足人物、地点、时代、事件或特定事实主题的数据查询需求、人物或实体之间逻辑或关联关系的延伸查询需求、时空主题范围的统计分析需求以及可视化呈现的需求等等。现在的认知计算技术结合了机器学习和人工智能，已经能够提供语词概念或图像实体的提取与分析、特征比较、相似性聚类等，数字人文平台完全可以应用最新技术，实现最新功能。从平台的角度来看，还有较大可提升空间。

參、平台需求与方案设计

当今时代已不再可能举全国之力穷天下收藏，兴建四库全书那样的项目，开发包罗万象的知识平台，现在甚至连某一学科或主题领域的资源都不可能一网打尽。因此我们在构建数字人文平台或开发人文数据数据库时首先应考虑自身的优势和特点，选取一定的文献类型或学科主题，充分考虑服务对象特点和需求，设定有限目标，并做好长期建设的准备。

目前很多汉学资源收藏机构已经开发了一些颇具特色的数字人文平台，如 CBDB、Docusky、MARKUS 等，应用了许多先进理念和最新技术。本文并非简单地想介绍上海图书馆的想法和平台，而是希望着眼于未来互联互通，对构建一个整体化的汉学数字人文研究的基础设施提出一些设想。包弼德教授曾在 2018 年提出过类似的想法，他建议构建一个“中国研究的网络基础设施”，希望通过各国汉学资料收藏机构的密切合作，开发一个通用平

台[7]，使汉学资源能够互联互通，进一步促成共建共享。

这是一个非常有远见、有现实意义和可行的建议，但这个平台不必是“一个”平台，而可以是整个汉学基础设施共同构成的分布式网络，即可以由相关汉学资源收藏和研究机构各自建设，但遵循共同制订的技术标准和互操作协议，这样就保证了资源获取和服务的互联互通；同时制订一定的合作机制和业务模式，这样又能够促进互惠互利和可持续发展。

因此，本文探讨的平台即是一个在功能上力求完善、能够满足当下研究需求的独立的数字人文平台，又同时在体系架构上兼顾了基于最新语义互联网技术的互联互通，是一个尚未实现但完全具有可操作性的设计方案（如图 2 所示）。

图 2 数字人文平台需求设计

一、平台的应用系统

平台的应用系统需具有一定先进性，这主要从两个维度来考察：系统维度和应用维度，系统维度主要指系统架构的先进性，又可以分为技术架构和内容架构，技术架构提供功能实现，内容架构提供研究视图，共同满足数字人文研究需求，解决数字人文研究的痛点问题；应用维度是指是否能提供应用机构完整全面的解决方案，包括馆藏业务系统、数字保存 / 典藏系统、知识库系统、服务展示 / 应用系统等。

7. 包弼德等 . 数字人文与中国研究的网络基础设施建设 [J]. 圖書館雜誌 2018.11.18-25.

图 3 数字人文平台系统需求

(一) 系统维度

1. 技术架构

系统纬度首先看技术架构。目前以微服务、容器、容器编排、服务网格、开发运维一体化（DevOps）、无服务器架构等理念为特征的新一代“云原生”技术正在席卷互联网应用。它拥有传统 IT 无法比拟的优势，可以帮助用户高效享受云技术的弹性和灵活性，使应用进一步微型化、轻型化，支持更加灵活的松散耦合，更加独立于底层基础设施平台，从而能实现热插入、平滑迁移、快速开发、迅速扩展、稳定运维、高容错等，大大降低应用成本，提高运行效率。目前云原生已经成为云时代新的技术标准。

当前还没有数字人文机构采用云原生技术，但图书馆界正在引入基于微服务的新一代图书馆服务平台，即 FOLIO，这是一个开源平台，可望采用云原生技术进行部署实施，其前后台分离的设计和“平台 +App”的架构有助于形成一个开放的软件应用生态，且设计理所当然地支持眼下另一个如日中天的技术概念，即“中台”，独立的数据中台（包括 AI 中台）将支持数字人文的各类应用模块灵活调用，在技术架构上具有无可比拟的先进性。当然该技术毕竟发展还不到十年，其成熟度和标准化程度还不是太高，微服务带来的应用复杂性还难以预料和掌控，这也是新技术必然带来的风险。

从数字人文的应用场景来看，有这样一些需求涉及技术架构：

（1）知识单元的标识及其管理问题。所有对人文研究具有独立意义的实体或信

息单元，如文献，或人、地、时、事、物、事件、概念，以及各类属性和取值词表等，都需要有独立的标识（即 ID），并统一 ID 编码标准，通常用 HTTP URI，其相互之间的关系如有必要可以通过建立本体知识库来管理。当然建立过程可以采用自动抽取加人工辅助校验方式。

（2）支持跨网域的搜索发现或链接功能。例如 OAI-PMH 规范，各类 Restful+JSON 的 API 规范、联邦检索页面分析规范等等。

（3）微服务的容器及编排规范。

（4）底层关系数据库和图数据库对数据操作的管理。

（5）用户及权限管理。

2. 内容架构

内容架构是数字人文应用系统非常独特的架构，也是语义技术逐渐成熟带来的一种能力，具体指数位人文平台中的数字化知识内容是具有一定结构的，这个结构可以以知识本体、关联数据、知识图谱等方式进行描述和表达，例如以各类描述词表对人物、地点、时间、事件和各类对象的各类属性和关系进行编码，使计算器可以对表达知识的这些语义数据（可以理解为 RDF 数据）进行操作，从而可以认为这些数据是机器可“理解”的，以至于可以认为整个知识库中的大量内容都是真实世界的一种映像，甚至可以能够让机器进行一定的“事实推理”。传统的数据库只能对字符串或二进制数据（如图像数据）进行操控，如全文检索也就是一种完全基于字符的匹配。数字人文平台对于信息资源的描述和组织可以认为是一种“数据化”过程，这一过程不一定完全依靠人类来做，很多都可以通过目前越来越成熟的机器学习和人工智能来实现。一旦机器能够读“懂”存储的信息所蕴含的知识内容，数字人文平台就能帮人文学者做很多事情，可以成为能力超强的“研究助理”，它不会遗忘任何一个知识细节，并且具有超快的计算能力。

有这样一些需求涉及内容架构：

（1）一致性 / 相似性计算。

（2）工作流定义对研究流程的支持。

（3）各类图像功能（如图像查询、对比、标注等）的支持。

（4）文本与图像关联（可提供加工平台，或研究对比）。

（5）提供证据链服务（纪录从底层文献到研究结果的整个过程中实体来源及变化，包括引用参考等）。

（6）海量数据可视化支持（遥读）。

（7）事实的可信度计算及排序（需建立可迭代的可信度模型）。

（8）众包数据加工平台的数据管理。

（9）数据系统迭代进化的支持（数字化、文本化、数据化）。

(二) 应用维度

数字人文平台大多由人类记忆机构，如图书馆、博物馆、美术馆、档案馆等进行建设和维护。作为数字人文基础设施的主要组成机构，他们的主要业务和服务都是围绕人文资源展开的，一个较为完整的平台通常分为四个层次：

1. 馆藏业务管理系统

这主要指对物理藏品或数字藏品的载体，从收集、入藏到转移、剔除或损毁的整个生命周期过程的管理，包括藏品管理系统。它提供了所有馆藏内容最初的来源和版本信息，是循证研究的源头，并通过业务过程的管理保证整个馆藏体系是一个不断发展变化的“活”的有机体。

2. 长期保存 / 典藏系统

即上述业务管理系统中的藏品管理系统的数字化版本，通常是能够保留最真实和完整信息的保存级数位文件，借助显示或其它设备，能够还原物理藏品的内容或型态，高级形式可以看成是每个馆藏的“数字孪生”，可供研究人员进行各种实验、模拟和深度研究。当然任何数字化版本都不可能保留原始对象的所有信息，总是会有所损失，所以依赖技术的不断进步，未来可能需要对馆藏进行再次数字化。这类系统目前主要采用关系型数据库加文件系统的方式实现，更为先进的采用了 NoSQL 数据库的大数据方式，基于云服务架构，而现在应该采用云原生架构加数据中台方式，这样就能够提供底层藏品管理系统与上层知识库系统之间的桥梁，同时提供大量的 API 供知识库系统和服务应用展示前台调用[8.]，这些 API 可以以标准方式发布于互联网，从而实现数字人文平台的全网域互操作。鉴于将来的数字人文研究都是基于数据的研究，有了这样的典藏系统，就可以解决绝大多数人文学者研究与教学的需要。

3. 知识库系统

目前似乎还没有一个恰当的术语来描述这样一种系统，最接近的词汇可能就是“语义知识库系统”，指应用了语义万维网技术对领域知识建立相互关联的知识体系，其知识单元是采用 RDF 形式（即主 - 谓 - 宾结构）描述的语义判断，而整个知识大厦是用知识本体语言 OWL 或 OWL2 组织起来，其背后的数学基础是一元谓词逻辑。数字人文平台的内容架构主要是由知识库系统提供的。其简化版就

8. 鲁丹、李欣、陈金传 . 基于 API 技术的数字人文基础设施的构建 [J]. 图书馆学研究 , 2019(13):42-46,57.

是采用关联数据的系统，更简化的一个版本是目前十分热门的利用“知识图谱”技术所支持的系统。这类系统在人工智能领域属于“符号学派”，与过去的专家系统同属一类，是将人的知识代码化形成规模之后，就具备了某种智能，现在与连结学派和概率学派有融合的趋势，作为人工标注或结构化的数据提供机器学习，从而具有自动的知识获取能力。数字人文平台需要大量的底层“知识库”来支撑各类数据的语义解释和关联关系，例如人名、地名、机构名、朝代、官职、谱系、辞典、词表等等，几乎所有的工具书都可以提供知识关联，所有的知识生产都是建立在过去知识的基础上，与这些底层知识库都可以建立起逻辑联系，最强大的是这些知识库都是以某种方式在整个互联网上提供共享，所有基于知识库和标准描述方式的术语词表都可以达成全网域的语义互操作。

4. 服务应用展示系统

这是数字人文平台中绝大多数功能得以实现和展现的前台，也是各类工具与后台数据进行链接的中介，通常以桌面或移动应用，以及浏览器方式提供。所有的搜索、浏览、展示（包括可视化）、众包和用户空间功能都在这里以 App 方式提供，这样有助于达成大量的第三方应用 app 的开发和发布，行成一个开放强大的数字人文应用和工具的生态环境，从而很容易实现包弼德教授提出的、为第三方数据、第三方工具、第三方图书馆免费公开的元数据访问和数据共享的规范和方案[9.]。

二、平台的资源、功能与界面需求

资源完整、功能完备、界面友好，是任何一个信息系统的基本要求。当然不同的系统对这三个方面的具体需求是不同的。一个好的数字人文平台至少要在这三个方面达到最低要求，同时要注意三者之间的平衡。

(一) 资源完整性

人文研究者在选定了研究问题之后，第一步就是要查询数据。很多机构在建设数据库或提供查询时只从自己已有的或订购的资源入手，这是不够的，还必须考虑到是否有办法提供外部资源的发现，甚至直接获取。要实现这一点，就要应用元数据收割方案，例如 OAI-PMH，或开发标准或个性化的 API，其中涉及很多考虑因素和资源互操作的具体技术，包括利用知识库系统实现不同系统间的语义互操作。如下图所示。

9. 包弼德等 . 数字人文与中国研究的网络基础设施建设 [J]. 圖書館雜誌 2018.11.18-25.

图 4 数字人文平台资源需求

(二) 功能完备性

数字人文平台需要考虑很多与过去数据库检索系统不同的功能，过去的系统主要是以文献为主要内容，根据数据库字段（即高级检索）或全文检索能够定位到具体的文献，再通过链接解析或其他方式获得原文。而数字人文系统由于提供了以“数据”为基础的存储、关联和查询能力，因此多了与“知识库“相关的很多语义功能，而且在搜索、浏览、管理等方面都能够全面支持基于知识的操作（例如 SPARQL 查询、分面组配等），有时甚至有包含逻辑推理的功能实现（如启发式搜索）。

数字人文平台还有一个特质是要利用众包让用户参与到系统的建设中来，这是当前几乎所有数字人文应用都采取的方式，因为仅仅通过图书馆或相关机构工作人员的工作是不可能实现海量高质量数据加工的。

图 5 数字人文平台功能需求

(三) 用户友好性

当前的信息系统对用户友好性的要求越来越高，这也是对系统界面提出的要求，除了一般的方便友好、美观简洁之外，能否提供良好的个性化服务成为系统是否具有粘性、能否留住用户的重要特性，而且个性化服务大量采用了人工智能技术，详见下图。当然，由于个性化的前提是需要有用户注册登录等用户管理功能，且对用户的行为也会进行一定的收集，这涉及到用户隐私问题，平台在设计开发时必须考虑到隐私保护与个性化之间的平衡，很多研究工具的提供应该能同时支持本地脱机版和上传网络版两种不同的运行方式，当然两者在功能细节上可以有所不同。

图 6 数字人文平台界面需求

三、平台的工具开发

利用大量的数字人文工具进行研究是数字人文区别于传统人文最重要的特点之一。工具是方法的重要组成，成熟的方法往往通过工具的开发而得以固化，并且负载了大量前人的经验总结。传统人文研究能够独立的工具不多，且数据的收集、阅读和加工处理往往是一体化、个人化的，工具很难独立于资料，有的甚至很难独立于研究团队。这也是为什么有许多人文社会科学学派往往是得益于独特的方法。

工具要求越丰富越好，但这里讨论的只是人文研究可能用到的具有一定通用性的工具，以及这些工具的常见功能，数字人文学者可以通过这些工具的组合，结合资源和研究过程，发展出自己独特的方法。这些工具可以有一定的独立性，但依附于平台能够更好地发挥作用，因此平台将致力于深入研究人文学者的需求，推出大量的标准规范，从而让大量第三方都能够开发自己的独特工具，甚至工具与资源或知识库的结合体，从而有助于形成一个应用生态，以及工具 App 市场。

这里将工具划分为平台性工具（包括数据工具、IIIF、GIS、文献计量工具、阅读工具、社会关系工具）、文本工具、图像工具、知识图谱工具、机器学习工具和可视化工具六大类。上述分类的合理性需要进一步探讨，其中涉及的内容也远不是对各类工具的穷尽例举，仅仅作为一个讨论的基础，供具体进行工具开发和平台建设参考。

图 7 数字人文平台工具需求

(一) 平台性工具

这里的平台是指网络上可以实现一定的功能、有特定输入输出的环境，平台性工具就是依附于平台或自身就可以独立成为一个平台的工具，它通常不是一个简单工具，而需要结合一定的数据，需要一些组件的配合，并经过一定的流程才能达到目的。例如 IIIF（国际图像互操作框架）就是一个功能强大的综合性图片平台，由多个服务器灵活组合而成，它本身就可以成为数字人文的服务平台，这里之所以作为一种工具，因为它提供了大量的关于图像的操作功能，如搜索、缩放、旋转、标注、比较等等，可以应用于人文研究，非常强大。类似的还有数据处理平台工具、GIS 平台工具、文献计量平台工具、社会网络分析工具以及阅读平台工具等等。

(二) 文本工具

文本是数字人文利用最多的资源类型，文本工具也是数字人文工具中种类最多、使用最频繁的工具，也是目前开发最成熟的工具类型。上图列出的是常用工具，一些综合性的文本工具，如“远读”、“细读”列在平台性工具类目下。

(三) 图像工具

通常所有的图像扫描、处理软件都可以作为数字人文的图像工具，这里仅列出数字人文项目非常常用的工具类型，如图像特征提取工具、图像分类 / 聚类工具和基于图像的搜索工具等，图像平台 IIIF 已作为平台类工具栏出。

(四) 知识图谱工具

知识图谱是数字典藏向数字人文进化的关键技术之一，这里将关联数据、语义万维网技术都归入知识图谱。这类工具包括了实体提取、URI 赋值、词表模式、本体构建等语义化工具，本体 / 词表管理、语义映像、RDF 语义数据存储等语义管理工具以及 SPARQL、启发式搜索、分面呈现等语义搜索、展示和利用工具等。

(五) 机器学习工具

当前数字人文的大量应用都用到了人工智能领域的机器学习技术。从 OCR 到实体提取，从神经网络到深度学习，无一不能应用于数字人文研究的各个过程。机器学习最大的特点是离不开数据，尤其是海量的数据，因此数字人文平台中的数据是其产生作用的前提条件，而由数据训练出来的机器学习模型又可以应用于更广泛的数据中，这是它的运作方式，也是它的价值所在。

(六) 可视化工具

可视化是数字人文进行数据操控、展示和结果呈现必不可少的工具，也是数字人文区别于传统人文的重要特质。可视化虽然有很多工具，但现在基于互联网的工具已成为主流，正在成熟起来。它后台连接的数据可以是平台上已有的数据，或者挖掘出来的数据，或者是用户上载的数据，是否支持多种应用方式取决于平台架构设计的灵活性。

肆、案例：历史人文大数据平台

上海图书馆正在建设的历史人文大数据平台，就是应用上述理念和技术，依托自身资源，向全社会提供一个先进、开放、全面的数字人文服务平台。打造这个平台主要有三个目的：一是升级原有的数字典藏系统；二是提供基于“知识”的数字人文服务；三是试验一些互联互通共建共享的新协议新模式。其实就是作为对前述数字人文发展趋势进行应对的一种尝试。

实现这三个目的有两条现实可行的路径：其一，从现有的数字典藏系统出发，也就是从目前上海图书馆馆藏特色资源出发，升级技术架构和内容架构：技术架构全面微服务化、容器化和平台化，支持外部资源与服务通过各种标准或非标准方式（推荐 Restful API）接入；内容架构进行“数据化”改造，支持“基于知识的服务”。其二，从数字人文研究者的角度出发，规划所有人文资源的整合方案，从提供资源、到提供平台环境（包括工具），努力实现主要数字人文应用场景的“一站式”的服务。

一、历史人文大数据平台的建设规划

上海图书馆走上数字化道路已经有四分之一个世纪。从 1996 年位于上海淮海中路的“新馆”开馆，就开始古籍数字化项目，并且参与了中国最早的由国家图书馆牵头的“试验性数字图书馆计划”，成立专门部门，每年耗费巨资进行特色资源的数字化工作，从无间断。

仅仅数字化是不够的，提供知识服务是图书馆的根本宗旨。早期重视数字化，但对于数字典藏系统的建设并没有充分重视，因此数字资源的整合服务一直没有充分开展。到 2016 年，上海图书馆尝试以最具特色的馆藏家谱资源为案例，开始了以服务为导向的系统开发尝试，取得了不错的效果，迄今家谱系统一直是数字典藏中利用效果最好的资源之一。

为了建设具有知识关联的数字人文服务系统，底层知识库平台建设是必不可少的，这也

是数字人文基础设施最困难的内容。近几年我们陆续构建了人名规范、地名规范、地理名称规范、机构规范等等规范知识库，可以支持目前列入计划的特色资源库的底层知识关联，并开始开发一些工具，提供众包、标注、分析、可视化等。

正是由于有了底层知识库的支持，上海图书馆的特色资源库才有可能做一个全面规划，将来各类数字人文系统可以在一个统一的平台上，我们称之为历史人文大数据平台。虽然这一平台尚未建成，但已经经过了初步尝试，证明了技术和工程上的可行性和可能性，且数据也有一定规模。目前我们除家谱库外，正在开发的还有古籍库（包括精品善本库）、碑帖库、地方志库、手稿尺牍库、名人档案库（如盛宣怀档案、张佩纶档案等）、民国资源库（包括书刊报）等，这些文献如按照数字人文研究的要求，可以建立无数个基于各类学科或主题的知识库，可以汇总在一个平台上提供满足各类需求的统一服务，通过一定的开放链接协议，可以将全网域的各类资源连为一体，组成一个虚拟汉学数字人文平台。

二、平台应用场景

对于一个资源众多、用户复杂、目标并不单一的服务平台来说，“主页”概念是不适用的。历史人文大数据平台虽然设计了一个主入口（主页，如图），但它的作用只相当于“游客中心”甚至是“疏散中心”，主要起到宣传、导航、资源发现和用户培训的作用。任何一个简单的搜索，都可以返回所有资源库中（甚至外部联邦检索或搜索引擎）的命中内容，这样能够让“随便逛逛”的读者也有所收获，同时用户对自己感兴趣的主题可以通过哪些资源库获得有一个非常直观的认识，从而带有目的的读者能够迅速找到属于自己的入口。

图 8 ：历史人文大数据平台主入口

平台对所有的专题库（包括文献库、知识库和工具库三类）都有一个入口，有一个页面提供了一个完整列表（如下图）。其中大多数文献库都是元数据库加扫描图片方式提供，个别有全文，知识库和工具库都支持响应式 H5 接口（可嵌入各类 App）。

图 9 ：平台各类资源库和功能入口

我们把平台用户分为四类：普通用户、专业用户、系统用户和机器用户。普通用户是无需用户认证即可来“随便逛逛”的用户，平台会有很多在线展览、人文讲座、推广活动、技能培训等内容发布。专业用户是平台服务的主体，通常是经过注册的研究人员或大学师生，也可能是相关机构中的个人用户（登录为单位用户或以 ip 控制方式提供权限管理），这类用户除非使用主页中的搜索框进行资源发现（搜索框在各相关页面也都会出现），一般无须从主入口进入，只要浏览器保留了登录 Cookie，域名会直接将其定位到他自己的个性化页面，该页面已经将其经常使用或可能用到的专业资源入口与各类服务功能集成在一起了，每个用户的专业入口都是个性化的，与“我的空间”捆绑，用户如果不满意，也可以在“我的空间”中修改参数设定。系统用户是那种参与数据加工或项目研发的用户，这是“平台性”的具体体现，作为平台，不是一个私有的封闭系统，而应该有一定的开放性，属于整个小区，允许大家参与共建、分享成果，因此必然有一类用户通过贡献内容、参与数据加工或功能开发而具有更多的权限。最后一类“机器用户”就是指通过 API 或其他接口直接消费数据的计算器程序，这样能将平台与互联网上其他应用连为一体，使“一站式”服务成为可能。

平台提供的所有服务可以分为“场景”、“故事”和“功能”三个层次，区别三类用户，提供不同的功能组合，详细如下表所示。“场景”可粗略地对应于数字人文研究的“行为”，例如搜索、浏览、下载、阅读等等；“故事”是组成场景的若干种应用；而“功能”是平台提供的最小单位模块，通常对应于目前云原生应用架构中的“微服务”。这里的服务基本都是用户直接可用的部分内容，后台其实还有大量的微服务，由于与平台用户并无直接关系，这里就不详述了。

表 1 历史人文大数据平台主要功能一览

用户类别 / 场景类别		普通用户 (非注册)	专业用户 (注册)	系统用户
Epic	Story 故事	Feature 功能		
搜索	简单搜索 / 全文搜索	精确或模糊匹配可选		
		支持提问自动分词		
		支持前后缀搜索		
	高级搜索	分字段限定	支持专业检索，即逻辑表达式在一个搜索框内实现高级搜索	
			支持字段按数据类型（如年代）限定，支持范围限定	
			位置限定搜索	支持正则表达式搜索
				统计式搜索
		智慧搜索（框式搜索）：I feel lucky （用各种技术猜测用户需求的检索）		
		二次搜索：在任意结果集中再进行限定搜索（支持逻辑式限定）		
	知识搜索		基于实体名称 / 概念的搜索，如人名、地名、机构名、事件、物体	
			启发式搜索	
				SPARQL/CQL 搜索（部分）
	图像搜索		基于输入图像的匹配（只支持部分图像库）	
浏览	排序	检索结果支持多字段排序 (分主次)，支持拼音 / 笔画 / 时间等排序方式，支持正序 / 逆序		
	分面组配浏览导航	按结果集分面情况选择		
	知识导航		按知识本体呈现结果 (相关图式参见可视化部分)	
	地图浏览	检索结果具有地理或时代属性的，可以在地图上进行时空呈现		
	对比		不同检索结果集比较呈现，可多窗口（参数可选？）对象比较参加阅读场景	
	结果可视化	结果集基本属性的可视化	下载至个人空间，再支持本地下载	可视化参数可调
下载	元数据下载至个人空间		需经授权才能下载	可订制 Schema 数据格式，与高清对象数据一起打包下载
	对象数据下载		可选择一定的参考文献各式，批量下载	
	参考文献格式下载		可選擇一定的参考文獻各式，批量下載	
阅读	结构导览	目录章节导览、各类跳转、超链接		
	文内搜索	搜索词高亮并可导航（下一个）		
	辞典工具	可外挂多语种 / 专业辞典		
	文本朗读	自动机器语音朗读 / 生僻字词朗读（随辞典工具）		
	书签 / 高亮	非专业注册用户存于本地，专业用户上载个人空间，并提供社会化（分享）选项		
	批注 / 修改		以 W3C annotation 标准方式（RDF）实现，可分享	
	对比		选择不同对象多窗口打开比较和分别批注（元数据结果集比较参见浏览场景）	
	文献推荐		阅读及写作过程中不断推荐相关文献，推荐模型相关参数 / 阀值可选	

用户类别 / 场景类别		普通用户（非注册）	专业用户（注册）	系统用户
Epic	Story 故事	Feature 功能		
高级阅读 / 阅读工具	细读		文本标注：根据词表（内建 / 挂接词表或用户导入词表）标注	
			文本标注：句读、分词、词性标注等	
			实体识别并标注	
			文白翻译	
		文本互译（外挂翻译工具）		
			文本格式化，格式标注 / 转换 （如 TEI、ePub 等，参见数据加工工具）	
	遥读	特征词 / 词云生成		
			自动摘要（模型参数可调）	
			文本 / 图像相似性计算、聚类分析	
			风格分析	
			情感分析	
	共读		支持用户书签 / 标注 / 批注等信息的共享和挖掘	
众包平台	任务管理			任务发布、分配
	用户管理			注册及权限管理、资金监管
	工作认领			工作量计算、统计反馈
	质量管理			质量校验、纠错、版本控制
可视化	文本可视化	各种词云	各种高阶词云、关系图、热力图	参数可调
	关系可视化		各类 d3/eCharts 图示	参数可调
	GIS 时空	行迹图	时空叙事 / 现地研究	
	图谱可视化		各类知识本体可视化，包括人物生平 / 大事年表 / 互动展览 / 组织机构图标等	

伍、结论

数字人文平台建设的愿景是让人文研究不再困难。从雅典学园到文艺复兴，从鲁国杏坛到康梁变法，两千年来人文学者的创造性思考从来都是依靠个体的博览群书与博闻强记，依靠师徒私授或学派论战，思想的诞生、学说的完善，以及对社会实践的影响主要依靠的是个人的能力，人文知识的产生、发展和传播的整个过程是偶然、不清晰和不确定的，每位学者都要从最原始的篇章学起，遍历所有典籍并考察整个源流，穷极一生只能成为专家而无法成就大家，而数字人文正在第一次给人文研究带来革命。

针对人文研究的完整过程，数字人文已能够分而治之：首先使数据查询和获取不再困难，其次使知识存储、传播和利用不再困难，然后让分析、比较，形成观点不再困难，最后

使结果展示、交流和争鸣不再困难，人文学者不再是单打独斗而是集团作战，无须管中窥豹而是直接综揽全局尽情把握，人文研究的规律与方法将得到更好的揭示，人文成果的发表形式将不再是报刊杂志，人文学说的比较与评价将更方便地在实践中得到检验和反馈，为人文研究提供的服务能力将更快地得到迭代和提高。

照此发展下去，那么问题来了：如果数字人文充分采用了人工智能技术，推向极致，可能机器也能自动进行人文研究。此时的人文，还是人文吗？

其实数字人文的终极意义还是在于以科技强化人文，而不是将人文变成被动机械的对象，进行去价值化和无意义化。最终的意义呈现，其主体是人类自身。当所有的人文都是数字人文时，“数字”与“人文”才能够真正合为一体，那时数字的工具性特征便不再重要，人文研究此时便能回归本源，真正彰显人类的价值和生命的意义。

这也是我们要用尽所有先进技术，推进数字人文平台的开发与建设的根本原因所在。

参考文献

[1] 尤西林 . 阐释并守护世界意义的人—人文知识分子的起源及其使命 [M]. 上海：华东师范大学出版社，2017：1.

[2] Tony Hey 等着，潘教峰等译 . 第四范式：数据密集型科学发现 [M]. 北京：科学出版社，2012：i.

[3] McCarty W. Humanity Computing. [EB/OL].

http://www.mccarty.org.uk/essays/McCarty,%20Humanities%20computing.pdf

[4] 刘炜等 . 面向人文研究的国家数据基础设施建设 [J]. 中国图书馆学报 2016.5.29-39.

[5] 包弼德等 . 数字人文与中国研究的网络基础设施建设 [J]. 图书馆杂志 2018.11.18-25.

[6] 王晓光“新技术”和“文科”不能简单相加 光明日报 2020.12.29.p14. https://epaper.gmw.cn/gmrb/html/2020-12/29/nw.D110000gmrb_20201229_3-14.htm.

[7] 包弼德等 . 数字人文与中国研究的网络基础设施建设 [J]. 图书馆杂志 2018.11.18-25.

[8] 鲁丹、李欣、陈金传 . 基于 API 技术的数字人文基础设施的构建 [J]. 图书馆学研究 , 2019(13):42-46,57.

[9] 包弼德等 . 数字人文与中国研究的网络基础设施建设 [J]. 图书馆杂志 2018.11.18-25.