LIBRA技术理论及其在史料图像资源中的应用*

□陈涛　李惠　张永娟　孙安

摘要　数字人文致力于数字技术与人文研究的深度融合，新技术的不断涌现，推动着数字人文的发展与变革。文章从众多的数字技术中凝练出五类最常用的关键性技术，形成LIBRA技术体系，主要包括关联数据(Linked Data)、国际图像互操作框架(IIIF)、大数据(Big Data)、资源描述框架(RDF)和人工智能(AI)。其中，资源描述框架和关联数据作为语义网的核心技术已在数字人文领域盛行多年；国际图像互操作框架作为新兴的技术已成为文化遗产机构开展图像资源研究的主流方案；大数据和人工智能技术已引起了人类社会的剧变，强大的算力和智能的算法赋予人文研究新的范式和新的视野。可以说，LIBRA已成为或将成为数字人文建设的核心技术框架，并会得到越来越多的重视和应用。在案例部分，文章从知识组织模型、数据存储模型和动态文本识别三方面探讨了LIBRA技术在构建多维度图像智慧系统中的深度应用。随着技术的发展变化，LIBRA并非一成不变，新技术也会补充和重塑LIBRA体系，以助力中国数字人文的建设与腾飞。

关键词　LIBRA　数字人文　国际图像互操作框架　人工智能　多维度图像智慧系统
分类号　G254
DOI　10.16603/j.issn1002－1027.2022.04.009

1　引言

中华优秀文化兼收并蓄、博大精深，其中蕴含的思想观念、人文精神、道德规范等，给了中国人无穷无尽的滋养，深刻影响着当代中国人的精神世界，是我们在世界文化激荡中站稳脚跟的根基[1]。如何让文物“活”起来，让观众能够在“一眼千年”中感悟传统文化的深沉和厚重，是中国数字人文学者所应努力的方向和担当的责任之一。数字人文如今已经成为一个活跃的研究领域，吸引了越来越多的研究机构和学者参与到这个领域的研究中来。

数字人文研究呈现多样性、交叉性的特点，主要体现在：(1)研究资源多态，数字人文研究的资源是“超文本”，由图像、书籍、文献、乐谱、档案、手稿、音频、影像、实物等多种格式数据构成；(2)研究领域广阔，包含诸多值得关注的问题，如历史文献、古籍档案、文化遗产的数字化及数据化处理，民族民间文化的数字化记录与可视化呈现，基于计算机视觉分析的艺术图像分析与鉴定，面向人文问题的大数据分析等；(3)研究背景交叉，除了计算语言学、文学、哲学、历史学、考古学、地理学、图书情报、艺术学等传统人文领域的学者外，还可包含信息学、计算机技术、数字文化、媒体技术等领域的学者；(4)研究工具多样，数据采集工具、数据存储工具、可视化分析工具，时空分析工具，自然语言处理、文本分析、云计算、知识图谱、机器学习等。

近年来，新技术不断涌现，知识图谱、关联数据、大数据、5G通信网络技术、边缘计算、数据中台、GIS、3D、AR/VR/MR、区块链、量子科技等等，新技术的出现定会不断丰富和冲击着数字人文研究。数字信息技术的发展和应用，为人文科学的研究提供了新的方法和工具，丰富了人文科学研究的数据来源，拓展了人文科学研究的问题域，这无疑为人文科学的发展提供了新的机遇[2]。马费成教授指出，新技术为我国哲学社会科学研究带来了新的历史发展机遇，新场景、新视野、新方法、新工具的出现，使整个哲学社会科学的研究范式正在发生深刻变化[3]。

* 国家社会科学基金项目“数字人文中图像文本资源的语义化建设与开放图谱构建研究”(编号：19BTQ024)的研究成果之一。
通讯作者：陈涛，ORCID：0000-0002-6609-4914，邮箱：chent283@mail.sysu.edu.cn。

刘炜研究员构建的数字人文技术体系主要包括数字化技术、数据管理技术和数据分析技术、可视化技术、AR/VR 技术、机器学习技术等[4]。周庆山教授认为，当前数字人文领域需要重点关注如何运用大数据、人工智能、数字孪生等新技术实现人文资源的“活化”和再造[5]。

2 技术视野下的数字人文

数字人文之所以可以区别于传统的人文研究，主要是有了更多的学科交叉和更多的数字技术的引入。很多数字人文研究以数据驱动，也有一些数字人文研究可以归结于技术驱动，甚至可以说应用的技术一定程度上制约了数字人文研究所能达到的广度和深度。数字技术已广泛应用于人文研究中，如历史学学者借助 GIS 技术进行历史知识和历史事件的静态和动态的可视化展示研究[6]；考古学学者利用计算机和高光谱成像技术进行了 3D 虚拟遗址绘图、文物虚拟复原、色彩还原等[7]；文学学者通过研究文本中的代词分布窥探作者的情感[8]；语言学学者通过建立形式化的数学模型来分析和处理自然语言[9]。

表 1 中列出了数字人文领域影响力较大的一些项目，以及这些项目背后涉及到的主要数字技术应用。其特点为：(1)GIS 的应用较为广泛，如：历史地理信息系统(CHGIS)、数位人文学术研究平台(Docusky)、书信数字化工程、学术地图、高迁古村等众多项目都与地理信息相关；(2)文本分析、文本识别、机器学习也是常用的进行内容分析的主要方法，如：数位人文学术研究平台(Docusky)、边沁手稿[10]、历史人文大数据平台。很多研究中，文本分析作为了数据处理的中间过程，如：欧洲时光机[11]、书信数字化工程；(3)关联数据、知识图谱、本体等语义网技术也是数字人文常用的技术，如 CBDB 关联数据平台[12]、文化日本[13]、上海图书馆的历史人文大数据平台、董其昌大展、古籍数字化记忆再造工程、高迁古村等；(4)图像资源方面，国际图像互操作框架(IIIF)在众多技术中占有主导地位，盖蒂(Getty)博物馆、美国华盛顿国家艺术画廊、巴伐利亚州立图书馆、Biblissima 手稿库等只是众多 IIIF 应用中极少部分的代表。

除了表 1 中列出的项目外，武汉大学敦煌莫高窟多模态知识图谱采用了图像标注、知识图谱、关联数据、机器学习等相关技术[14]；北京大学宋代学术

表 1 国内外部分数字人文项目核心技术应用

项目名称	所属机构	主要技术
中国历代人物传记资料库(CBDB)、历史地理信息系统(CHGIS)	哈佛大学、台湾“中央研究院”、北京大学、复旦大学等	GIS、关联数据(上海图书馆版本 CBDB)、可视化(中文在线版本 CBDB)
数位人文学术研究平台(Docusky)	台湾大学	文本分析、GIS
边沁手稿	伦敦大学学院	文本识别
书信数字化工程	斯坦福大学	社会网络分析、GIS
文化日本	东京大学等	关联数据、IIIF、机器翻译、图数据库
欧洲时光机	TMO 协会	3D、机器学习、AR/VR
IIIF 应用(Getty 博物馆、美国华盛顿国家艺术画廊、巴伐利亚州立图书馆、Biblissima 手稿库等)	/	IIIF
历史人文大数据平台(家谱、古籍、盛档、报刊、近代图书、红色文献)	上海图书馆	关联数据、文本分析、聚类分析、数据中台、知识图谱、IIIF、图数据库
董其昌大展	上海博物馆	机器学习、知识图谱
学术地图	浙江大学、哈佛大学	GIS
古籍数字化记忆再造工程研究	武汉大学	关联数据、大数据、人工智能、知识图谱、人机交互、VR、IIIF
高迁古村	中国人民大学	3D、GIS、本体组织

传承语义网络使用了知识图谱、关联数据等技术[15]；华东师范大学数字人文研究支撑平台主要使用了国际图像互操作框架和关联数据技术[16]。同时，很多新技术也在逐渐渗透到数字人文研究领域，如中国博物馆协会的“博物馆在移动”项目汇集了130家国家一级博物馆，打造博物馆聚合平台，并在线上借助5G+技术，让观众与文物“亲密接触”、沉浸互动。

这些常用的数字技术中，GIS、知识图谱、社会网络分析常用于数据的呈现、分析；3D、AR/VR用于用户体验的提升；本体、关联数据、IIIF等技术常用于数据的组织；文本分析、图像标注、机器学习、聚类分析等常用于数据的处理。本文将这些数字技术进行归纳，形成LIBRA技术体系，以期对数字人文建设和研究提供技术方向的指导。

3 LIBRA与数字人文

图1为数字人文研究核心技术树状图，其中人文(Humanity)为“树根”，在这里多学科交叉、盘根错节，可见数字人文研究离不开人文的根基，需要人文精神、人文情怀；所有数据(Data)资源为“树叶”，树叶形态各异意为数据异构，树叶分布于不同的树枝则代表数据多源；各种数字人文研究成果则为“树果”。怎样构联起数字人文这棵大树，LIBRA给出了可行的实施方案和技术框架，LIBRA并不是某一种技术，而是数字人文基础设施建设中常用的五类技术总称。LIBRA主要包括：L－关联数据(Linked Data)、I－国际图像互操作框架(IIIF)、B－大数据(Big Data)、R－资源描述框架(RDF)和A－人工智能(AI)。其中，资源描述框架是“树干”，是人文与数据连接的主干道，是资源建设和应用的基础。关联数据和IIIF为“树枝”，他们在树干的基础上共同串联起了不同来源、不同类型的数据，让不同的数据个体智联成整体。其中关联数据主要针对文本型的结构化数据，而IIIF则主要应用于图像资源。大数据和人工智能技术是强有力的框架和工具，犹如剪刀一般修枝剪叶，增强光合作用，改变了树体的营养状况，这样数字人文大树才会枝繁叶茂、绿叶长青。

3.1 资源描述框架(RDF)

资源描述框架(Resource Description Framework，RDF)是一个使用XML语法来表示的数据模型，用来描述Web资源的特性以及资源与资源之

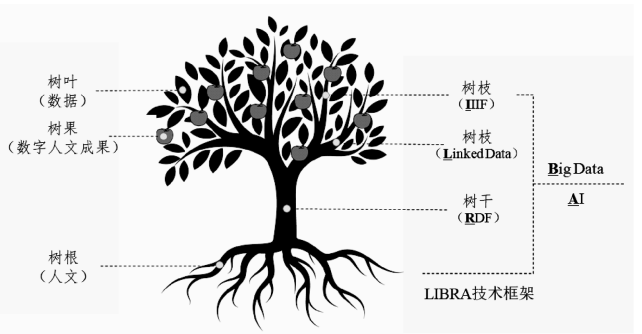


图1 数字人文研究核心技术树状图

间的关系。RDF主要用三元组(主、谓、宾三段式)来描述资源万物，由于其结构简单成为了语义网时代通用的数据交换形式和元数据模型，同时它也是知识图谱中常用的图模型之一(知识图谱中常采用属性图模型和RDF图模型)[17]。

数字人文研究资源是“超文本”，RDF在数字人文资源建设中的作用不言而喻，RDF为多源异构的数据资源提供了语法层面上的统一，使不同数据之间的融合成为了可能，也更为便捷。纵观国内外数字人文研究，多数研究机构和学者都将RDF作为资源组织的首选。从应用类型来看，本体设计、知识组织到实例数据发布都使用RDF来进行描述；从数据集规模的大小来看，小到单个本体文件和规范词表发布，大到数以亿条量级的知识库发布，以及各种特色专题库的数据组织也都采用了RDF。RDF数据的存储在工程应用中，建议使用图数据库(Graph DB)进行存储，三元组数据库(Triple Store或RDF Store)可以看成图数据库的一种类型，也得到较多应用。关系型数据库和三元组数据库对比见表2，从结构设计、调用方式、查询语言和运行效率方面进行了对比，并阐述了使用三元组数据库的优势所在。

表2 关系型数据库与三元组数据库的对比

对比项	关系型数据库	三元组数据库	优势
结构设计	需预先设计好表结构	无需事先制定数据结构	扩展容易 维护便捷
调用方式	数据库驱动连接	HTTP连接 (SPARQL端点)	随时随地 不受限制
查询语言	SQL语言	SPARQL语言	语言统一 迁移方便
运行效率	数据耦合性强，复杂查询效率高	低耦合，图路径查询效率高	海量数据 高并发性

3.2 关联数据(Linked Data)

关联数据近年来已成为数字人文研究，尤其是跨学科中多源异构资源整合的关键技术。需要注意到关联数据和数据关联并非同义，所有相关联的数据都可以看成是数据关联；而关联数据是语义网的轻量级实现，它不是新的数据，而是数据一种新的呈现形式。一般认为只有符合蒂姆·伯纳斯·李(Tim Berners－Lee)在2006年概述的关联数据的四个原则[18]，才被认为是关联数据。

图2显示了语义网七层框架结构和关联数据实现标准之间的对应关系，关联数据主要基于语义网七层框架的前四层进行展开，即实现用URI标识实体、用OWL组织实体、用RDF表述实体、用SPARQL检索实体。

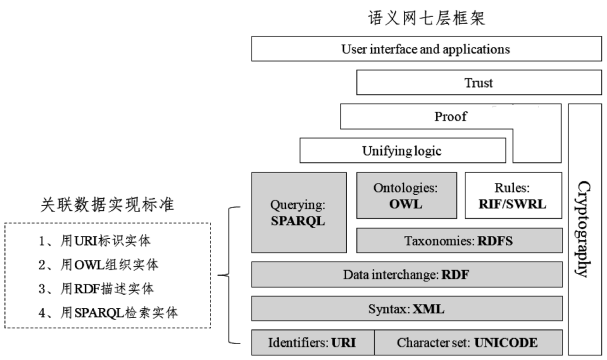


图2 关联数据实现标准

除此之外，关联数据还要求当资源被请求时，能够尽量提供与外部资源的链接，以便使用者获取更多的相关资源。常用的关联方式见表3，这里从“唯一码匹配”“属性值匹配”“图模式匹配”和“语义度匹配”四种匹配模式进行了说明。四种模式由易到难、由浅入深，在实际进行资源关联时，通常从最简单的关联模式开始，逐渐过渡到下一模式。关联好相应资源后，需要将匹配关系以三元组形式存储到资源RDF资源中，常用的资源关联属性有：owl:sameAs(链接相同资源)、foaf:homepage(链接到资源主页)、foaf:topic(链接到资源主题)、rdfs:seeAlso(链接到资源其他信息页)等，甚至所有的对象属性(owl:ObjectProperty)都可以作为资源之间的关联属性使用。

3.3 国际图像互操作框架(IIIF)

IIIF提供了一种前所未有的新方法，它是一组定义数字图书馆互操作性框架的标准，通过标准的

表3 数据源资源关联模式

关联模式	模式说明	模式举例
唯一码匹配	不同数据源资源如果有唯一标识码，可以使用该码进行匹配，相同即为同一资源	ISSN、DOI、ISBN等
属性值匹配	不同数据源资源拥有某个特殊的属性值，该值相同或达到某个相似程度即为同一资源，可结合字符相似度算法	机构名、经纬度信息、邮箱等
图模式匹配	不同数据源资源之间具有相同的子图(sub-graph)信息(多个属性具有相同值)	作者＋籍贯＋出生年月、文献名＋多作者等
语义度匹配	不同数据源资源之间的语义相同或相近，可结合机器学习方法对资源语义相似度进行判断	事件描述、机器翻译、人物传记等

应用程序编程接口(API)集，提供了一种在Web上描述、分发和访问图像的统一方法。该方法使用标准化的图像请求格式共享图像数字内容，提高了图像资源的在线研究能力。在众多机构的共同努力下发展起来的IIIF很快被更广泛的文化遗产部门所采用，在数字人文建设和研究中得到越来越多的关注。目前IIIF框架已推出的稳定版API有图像API(3.0版本)、呈现API(3.0版本)、认证API(1.0版本)和检索API(1.0版本)[19]。

目前不少机构对IIIF中的图像API和呈现API研究较多，也有大量的图像资源以IIIF要求和标准发布，而对图像的检索API研究，尤其是图像的语义关联研究很少涉及。图3描述了图像资源的语义关联流程，主要分为“内容标注和对象识别”“语义标注”“知识关联和知识发现”三步。其中，“内容标注和对象识别”主要对馆藏图像资料中的对象进行提取和注释，这里的“对象”可定义为图像中的任一实体或目标，如图像中的某个实体(人名、地名)、某个元素(花、鸟、树)等。对象的区域提取和内容注释一般采用人工标注的方式，对于一些有规则和图像质量较好的图像可以尝试使用机器学习的方式进行目标检测和自动标注。对象区域可以为矩形、圆形等规则区域或任意不规则形状区域，对象的每一条注释都将生成唯一的资源URI(资源主语)，并将注释内容以RDF三元组形式进行存储。“语义标注”实现了图像对象资源和外部关联数据集的关联，这里的关联关系为一条或多条RDF三元组。关联关系(谓语)可以使用已有本体中的对象属性，在

LOV 或者本体服务中心(OntHub)中可查询相关本体的对象属性。语义标注中的关联对象(宾语)为其他数据集或知识库中存在的资源 URI,而这些资源或多或少已经关联到其他的链接(开放)数据,从而实现更广范围的知识关联和知识发现。

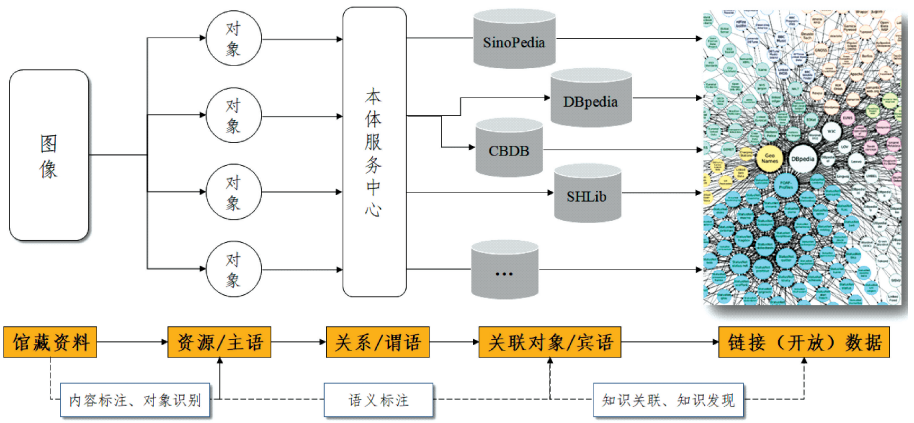


图 3　图像资源语义关联流程

3.4　大数据(Big Data)

大数据是信息技术发展的必然产物,更是信息化进程的新阶段,其发展推动了数字经济的形成与繁荣。数字人文研究更注重碎片化数据、海量数据、多源异构数据的采集、清洗、重组、分析与关联,进而深度揭示数据之间的内在关系。近年来,大数据获取、存储、管理、处理、分析等相关的技术已有显著进展,大大推动了数字人文研究的发展。

随着数字人文应用系统所涉及的数据量逐渐增大,不得不考虑采用大数据解决方案。大数据常提的 5V 特性有:规模大(Volume)、多样化(Variety)、高速性(Velocity)、价值化(Value)及准确性(Veracity)。同时,越来越多的大数据应用也引入语义技术,通过语义链接,给大数据系统带来开放性和互操作性,并提供基于"知识"的分析[20]。数字人文研究推崇开放、融合、智能,因此需要将大数据理念、关联数据思想和人工智能技术结合起来。图 4 显示了大数据 5V 特性的技术实现方案。

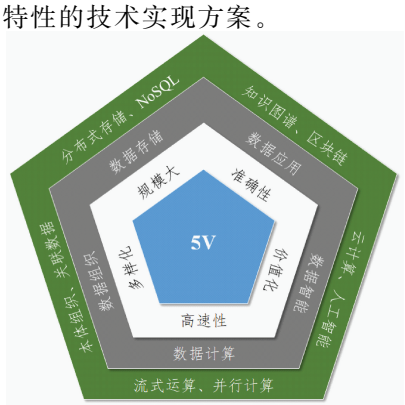


图 4　大数据 5V 特性技术实现

(1)规模大:数据存储的转变

大数据的采集、计算、存储量都非常庞大,如果还沿用传统的数据存储方式,必将给大数据分析和应用带来诸多不便。对于数量非常庞大的海量数据,目前的分布式数据库技术如 NoSQL 和 Hadoop 等都能很好地进行处理。文中讨论的三元组数据库(图数据库)同样也是 NoSQL 数据库的一种。图数据库特别适用于超大量的数据节点以一定的关系链接起来的形式,不管节点内部的数据多复杂,它都能高效地进行增删改查等操作。而且正由于它对节点内的数据没有限制,在进行大数据分析时往往更为高效,因而能得到更多的启发。

(2)多样化:数据组织的挑战

大数据的数据来自于多种数据源,且数据种类多样,已突破了以前所限定的结构化数据范畴,包含了大量的半结构化和非结构化数据。数据类型也不再局限于文本,还有图像、音频、视频、科学数据等多种类型的数据。本体、RDF、关联数据等技术的结合,为大数据提供了统一的知识组织模型、标准的数据交换方式和通用的资源融合模式,使多源异构数据的描述更为规范且富含机器可理解的语义,使大数据具有更好的开放性和互操作性,也将使大数据的分析深入到"知识"层次。

(3)高速性:数据计算的目标

大数据的数据增长速度快,因此要求获取数据和处理数据的速度也要快,需要从各种类型的数据中心快速获得高价值的信息。Spark 是一种混合式

的计算框架,是一种专为大规模数据处理而设计的快速通用的计算引擎,它自带实时流处理工具。此外 Storm、Samza 和 Flink 也是常见的流式框架。Spark 也可以与 Hadoop 集成代替 MapReduce 做并行计算。并行计算是增强复杂问题解决能力和提升性能的有效途径,其可以通过多种途径实现,包括多进程、多线程以及其他多种方式。

(4)价值化:数据智能的精髓

众所周知,大数据虽然拥有海量的信息,但是真正可用的数据可能只是很小的一部分,从海量的数据中挑选小部分数据工作量巨大,因此常将大数据分析和云计算联系起来。大数据必然无法用单台的计算机进行数据处理,必须依靠云计算灵活的张力和强大的算力。随着人工智能的快速应用及普及,深度学习及强化学习等算法不断优化,大数据技术将与人工智能技术更紧密地结合,具备对数据的理解、分析、发现和决策能力,从而能从数据中获取更准确、更深层次的知识,挖掘数据背后的价值,催生出新业态、新模式。

(5)准确性:数据应用的关键

数据的准确性和可信赖度,即数据的质量,是大数据发展的关键。关联数据由于其过度的开放性,一直被不少学者诟病其数据的质量问题。目前,完整性、准确性、一致性和及时性,常被用来作为评估数据质量好坏的指标。借助知识图谱和知识计算,对知识的可信度进行量化,通过舍弃置信度较低的知识来保障数据的质量。区块链的可追溯性使得数据采集、交易、流通,以及计算分析的每一步记录都可以留存在区块链上,使得数据的质量更加有保障。区块链技术的迅速崛起将有效突破大数据面临的困境,帮助大数据发挥更大的价值。

3.5 人工智能(AI)

随着信息技术的发展以及人工智能的出现,图书馆学、情报学领域开启了走向智能、智慧的演进和发展之路。数字人文对于人工智能的渴求尤为显著,目前,数字人文正在引领文化生产体系的数字转向,已经成为一个语言学、文学、史学、哲学、艺术学等传统人文学科与图书馆学情报学、计算机科学、人工智能等信息科学共同关注的新兴跨学科领域。就本文研究的 LIBRA 技术理论中,关联数据部分的资源关联匹配中的字符串相似度比较、语义度匹配,以及非结构化数据的实体抽取与识别等都离不开机器学习的算法。IIIF 中的图像对象轮廓提取、对象识别、图像自动标注等也离不开人工智能的相关算法,此外图像中基于深度学习的 OCR 技术也成为了行业主流。

人工智能在数字人文领域的应用中,有两类新兴的技术值得关注:知识图谱和 AI 中台。知识图谱是一种用图模型来描述知识和建模世界万物之间的关联关系的技术方法[21]。本质上,知识图谱是一种揭示实体之间关系的语义网络,它和 LIBRA 中的关联数据(L)有着千丝万缕的联系。大数据时代应运而生的中台得到越来越多的企业和机构的关注,数据中台、业务中台、算法中台、技术中台等接连涌现,“中台”概念的引入将对数字人文基础设施建设起到变革作用。

在众多中台中,有两类中台尤为值得关注:数据层面的数据中台和算法层面的 AI 中台,AI 中台用来连接业务中台和数据中台。数据中台可以实现多种功能,通过数据技术,对海量数据进行采集、计算、存储、加工、可视化,提供统一标准和口径。数据中台把数据统一之后,会形成标准数据,再进行存储,形成大数据资产层,进而为运营和管理提供高效服务。AI 中台通常由数据和算法组成,因此 AI 中台离不开数据中台,它是一个用来构建大规模智能服务的基础设施,对业务所需的算法模型提供了分步构建和全生命周期管理的服务,机构可以将自己的业务不断下沉为一个个算法模型,以达到快速复用、组合创新、规模化构建智能服务的目的。

图 5 给出了数字人文研究中常用的 AI 中台框架,主要包含文本和图像资源这两类数字人文研究中最常用的资料类型。文本资料经常需要进行自然语言处理方面的分析,如词频统计、语法句法的文本分析、命名实体识别、聚类分析、社会网络分析、字符相似度计算等;图像资料主要是针对图像的内容标注,包括图像目标(对象)检测、OCR 文本识别、机器自动标注、图像识别、图像处理等。而在知识图谱层面,可以进行数据挖掘、语义标注、知识关联和知识推理、知识计算等。当然这里仅列出了常用的 AI 功能部分,实际工程时可以根据需要增减功能模块,形成自己的 AI 中台架构。图中的 AI 功能在很多数字人文项目中都有所涉及,但多是和业务逻辑绑定,与具体平台耦合性太强,导致这些功能的可扩展和可复用成本太大。因此,在数字人文资源建设和基

础设施布局中,AI 中台的引入至关重要。

图 5　AI 中台框架

4　LIBRA 实践——多维度图像智慧系统

多维度图像智慧系统(Multi-Dimensional Image Smart System,MISS)[①]由 LIBRA 技术理论驱动,以图像为研究对象,实现古籍、文物、藏品等图像资源中的文字识别、图文识别、版本比对、特征提取、光谱分析等功能,并提供图像维护、发布、复用、标注等一站式服务,以达到数据驱动人文艺术研究创新的目的。MISS 平台作为数字人文图像资源建设和研究的典型案例,已得到上海图书馆、南京大学、上海交通大学、华东师范大学、上海大学等相关机构数字人文学者的肯定。现从"知识组织模型""知识存储模型""动态文本识别"几个方面探讨 LIBRA 在 MISS 中的实现。其中,知识组织模型部分主要采用 IIIF(I)框架对图像资源进行组织,并使用 RDF(R)进行图像组织、内容注释、语义关联等相关数据的描述,并使用关联数据(L)标准进行相关数据的发布和关联;知识存储模型体现了大数据(B)的存储理念,使用 NoSQL 进行高效存取;动态文本识别则结合机器学习等 AI(A)技术对图像中的文本进行动态 OCR,以提高交互体验。

4.1　知识组织模型

MISS 平台中最小单元为图像(Image),一幅或多幅图像通过画布(Canvas)组成一套藏品清单(Manifest),一套或多套藏品清单将组成藏品集合(Collection),而集合之间或集合和藏品清单能再次组合成上一层集合,依此类推可形成嵌套集合。集合对应的层次模型为:

集合 C={集合 C_0, … , 集合 C_m, … , 清单 M_1, … , 清单 M_n}

该模型表示一个集合 C 必须要有 1 个或多个清单 M,可有子集合 C_m。这里以"书画精品集"集合为例,用知识组织方式显示集合资源之间的关系。图 6 中 mc 的节点表示集合资源,mm 的节点表示藏品清单资源。

(1)"书画精品集(mc:C1)"下有 2 个子集"近现代

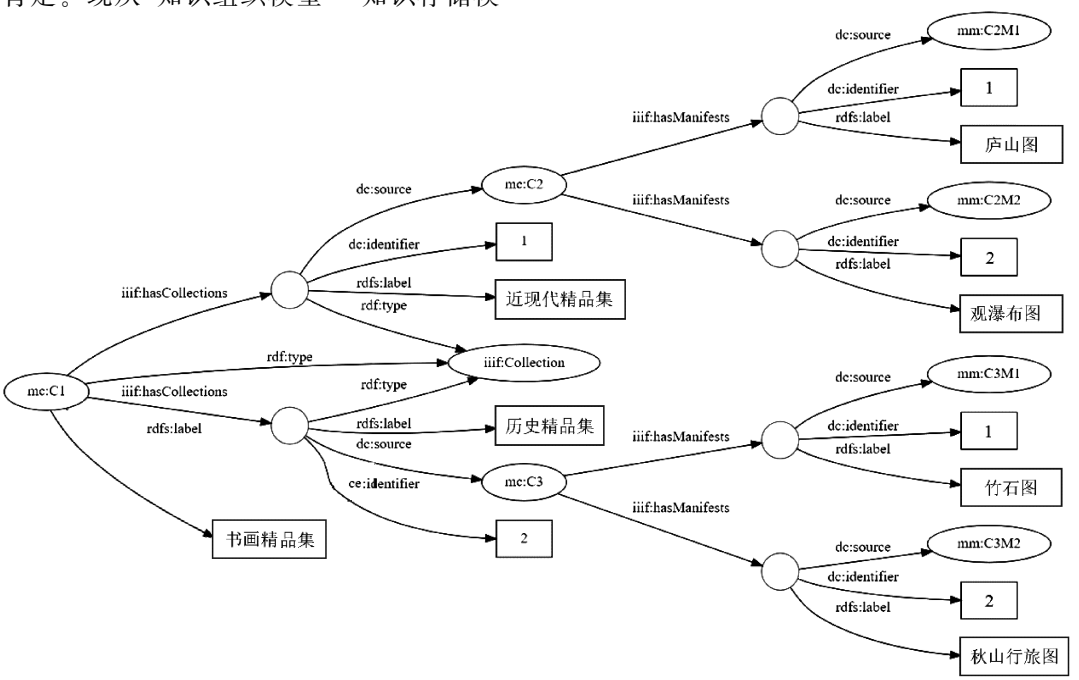


图 6　集合"书画精品集"知识组织

① 网址:http://miss.newwenke.com/sas.

精品集(mc:C2)”和“历史精品集(mc:C3)”,通过属性 iiif:hasCollections 连接三者类型都为 iiif:Collection,其中 mc:C2 的顺序为 1,mc:C3 的顺序为 2。

(2)集合 mc:C2 下 2 个藏品清单,分别为“庐山图(mm:C2M1)”和“观瀑布图(mm:C2M2)”,顺序为 1 和 2。集合和清单之间通过对象属性 iiif:hasManifests 相连接。

(3)集合 mc:C3 下同样含有 2 个藏品清单,依序为“竹石图(mm:C3M1)”“秋山行旅图(mm:C3M2)”。

图 7 显示了藏品“庐山图”的知识组织模型,这里可以详细看到 IIIF 框架中的 Presentation API(2.1 版本)的组织架构,主要有 iiif:Manifest、iiif:Sequence、iiif:Canvas、oa:Annotation 等四个核心类。

(1)藏品 mm:C2M1(庐山图)的类型为 iiif:Manifest(清单),并含有一些元数据属性,用 iiif:metadataLabels 进行赋值。

(2)iiif:Manifest 类下包含 iiif:Sequence(顺序)类,用来指定藏品的浏览顺序,用属性 iiif:hasSequences 连接。

(3)iiif:Sequence 类下含有 iiif:Canvas(画布)类,画布中包含了具体需要显示的图像,用属性 iiif:hasImageAnnotations 连接,示例中的图像为 img:lushantu,作为标注类(oa:Annotation)连接到画布 mm:C2M1-c1 中。

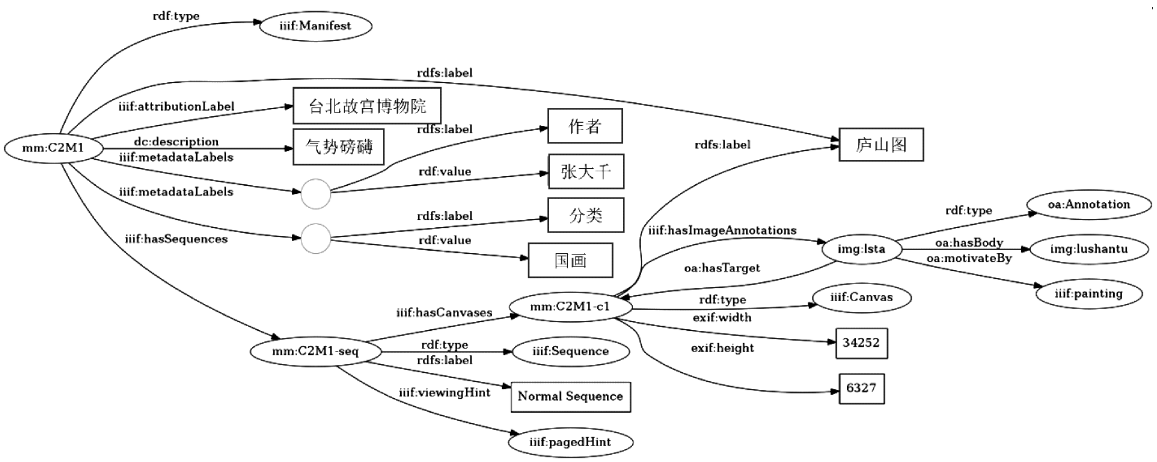


图 7　藏品“庐山图”知识组织

4.2　数据存储模型

MISS 平台的资源采用 NoSQL 数据库中的图数据库进行存储,存储时并不建议将所有的 RDF 数据都存于单一 Graph 中,图 8 显示了 MISS 平台的资源存储模型。模型中可以看出按照资源类型分为了四类 Graph:集合 Graph、清单 Graph、注释 Graph 和语义标注 Graph。

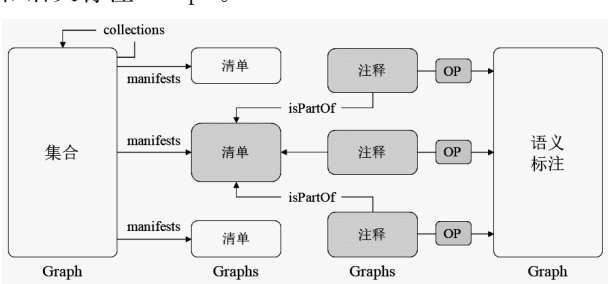


图 8　MISS 数据存储模型

(1)集合 Graph

用来存储所有集合信息,嵌套的集合也存在该 Graph 中。当某个集合含有清单链接时,将通过属性 iiif:hasManifests 链接到具体的清单 Graph,每个集合可以链接至多个清单 Graph(1:N)。依据关联数据的 URI 命名规则,可以将集合 Graph 定义为“{scheme}://{server}{/prefix}/graph/collections”。

(2)清单 Graph

用来存储具体的清单内容,每一个清单文件都将用独立的 Graph 进行存储。每个 Graph 中包含了清单藏品的 Metadata 信息、画布 Canvas 信息和图像 Image 信息。IIIF 要求每个清单文件保存为唯一的 JSON-LD 文件,并在网络中提供调用。因此在设计每个清单 Graph 的命名时,可以使用该文件的 HTTP 访问地址作为 URI 路径,即“{scheme}://{server}{/prefix}/manifest/{identifier}.json”。

(3)注释 Graph

类似于清单 Graph,也是每条注释都将存于独立的

Graph 中。该 Graph 中含有注释的具体内容,已经标注的图像方位;同时,也通过 dct:isPartOf 属性将该条注释指向具体的清单 Graph。每个清单 Graph 包含多条注释的 Graphs(1:N)。每条注释的 URI 不直接在平台中调用,因此可定义为"{scheme}://{server}{/prefix}/annotation/{identifier}"。

(4)语义标注 Graph

用来存储与每条注释相关的语义关联信息,这些语义关联信息将通过对象属性(OP)进行关联,并存储在单一 Graph 中。每条注释可以含有多条语义关联信息(1:N)。该 Graph 的 URI 为"{scheme}://{server}{/prefix}/graph/relation"。

4.3 动态文本识别

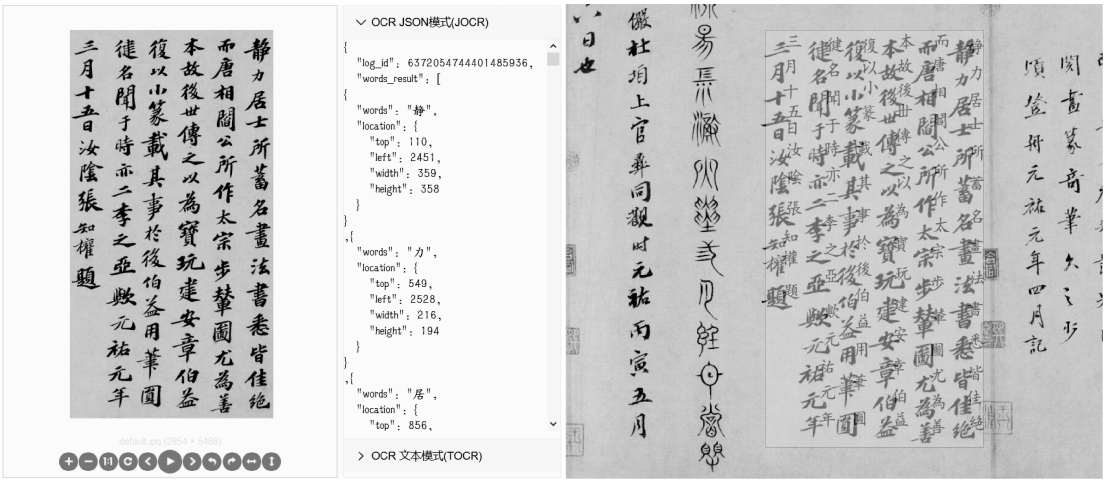
数字人文研究中经常需要对图像资源进行文本化处理,进而使用自然语言处理和文本挖掘等方法进行文本分析。机构在进行 OCR 时,经常遇到以下两点障碍:(1)馆藏机构具有大量有待 OCR 的数字资源,所有资源事先进行 OCR 识别成本太大;(2)对于某些尺寸较大的数字资源,事先进行 OCR 识别也不太现实,也很少有 OCR 厂商支持超大图像的文本识别。因此,如何将 OCR 环节从事前执行转移到事中运行,在研究中根据需要对资源进行实时动态 OCR 识别,是 MISS 平台的一次尝试。

借助 IIIF 框架中的图像 API 可以轻易地将需要识别的区域发送到 OCR 接口进行识别,识别模型主要使用图像处理和 LSTM 神经网络预测模型构建,训练步骤为:(1)图像预处理:对图像进行灰度、二值和降噪处理,形成黑字白底图像;(2)文本检测:采用分割的方法对图像中的文字进行分割,分割粒度为字符级,即分割成一个一个的方块字;(3)人工标注:对分割好的方块字进行人工标注;(4)训练学习:采用 LSTM 神经网络预测算法对人工标注语料进行学习,生成图像 OCR 识别引擎。

《步辇图》为唐代著名画家阎立本的名作,是唐代绘画的代表性作品,也是中国历史上最杰出的绘画作品之一。该画卷为绢本,纵 38.5cm,横 129cm,记载贞观十四年(公元 640 年)唐太宗召见吐蕃王国使臣禄东赞的场景,是唐朝经济文化强盛和古代汉族与藏族友好往来的历史见证,具有珍贵的历史研究和艺术价值。该画卷现收藏于北京故宫博物院,为中国十大传世名画之一。《步辇图》画卷有米芾等 22 位名人及收藏家的题跋,整张画卷有 2.2G(TIF 格式)大小,如此巨大的图像资源事先进行 OCR 极不现实,有些字体的识别效果也不理想。结合 IIIF 和构建的 OCR 识别模型,可以根据需要对相关题跋进行实时动态 OCR 识别。

动态文本识别时,使用 IIIF 的图像 API 对识别区域进行提取,图 9 为《步辇图》标注的三段题跋区域。以张知权的楷书题跋为例进行说明,这里的目标区域 URL 地址为"http://183.194.249.232:9002/iiif/yanlibenbuniantu.tif/37232,1217,2854,5488/full/0/default.jpg",通过图像 API 的试用,可以获取图像中任意区域。将该区域地址发送到 OCR 接口进行识别,识别结果如图 10(a)所示,在识别的 JSON 文件中,可以进行人工校正,以获得更高的准确度,图 10(b)为最终的结果呈现。

图 9 《步辇图》识别区域提取



(a)题跋在线 OCR 识别　　(b)识别结果呈现

图 10　《步辇图》实时 OCR 示例

5　总结与展望

数字人文研究需要采用大量的技术方法和技术手段，来实现科技和人文的跨界破壁。本文从众多的数字技术中提炼出对数字人文建设具有变革性的五类技术，即 LIBRA(关联数据、IIIF、大数据、RDF和人工智能)。在整个 LIBRA 体系中，资源描述框架(RDF)可用在数字人文研究中的资源描述部分，实现了异构数据间的语法统一；关联数据(Linked Data)更多的体现在数字人文相关数据的发布、共享和交互、融合方面，在多源数据之间建立起了语义关联链接；国际图像互操作框架(IIIF)主要针对数字人文研究中的图像资源，提供了不同机构间图像资源的可共享和互操作。大数据(Big Data)和人工智能(AI)并不单指某一种技术，它们是一类技术的总称，这两者的应用已经给各个领域都带来了剧变，分布式、云计算、自然语言处理、文本分析、机器学习等逐渐改变了人文研究的传统模式，推动着数字人文研究的发展和突变。

近年来，众多业界学者在研究和探索数字人文基础设施，总体来看，数字人文基础设施涉及网络基础设施、数据基础设施、技术基础设施、研究基础设施等多方面。LIBRA 将在技术基础设施部分发挥重要作用，从资源描述、知识组织、交互共享等方面提出了一定的通用标准和实施方案。LIBRA 中的五类技术在应用时，可根据应用的广度和研究的深度进行组合和扩展。当然，文中案例部分的 MISS 平台也仅是对 LIBRA 技术的粗浅尝试，技术也是处于不断发展变化之中，新技术也将会补充和重塑 LIBRA 技术体系。5G 通信网络技术、量子计算、区块链等新兴技术的发展终有一天会引入到数字人文研究中，必将带来数字人文发展翻天覆地的变化。

参考文献

1　曾军.数字人文的人文之维[N/OL].中国社会科学报，2020－08－28(7)[2021－03－10]. http://m.cssn.cn/zx/zx_bwyc/202008/t20200828_5175649.htm.

2　大卫·M·贝里，安德斯·费格约德.数字人文：数字时代的知识与批判[M].王晓光，译.大连：东北财经大学出版社有限公司，2019.

3　马费成，李志元.新文科背景下我国图书情报学科的发展前景[J].中国图书馆学报，2020，46(6)：4－15.

4　刘炜，叶鹰.数字人文的技术体系与理论结构探讨[J].中国图书馆学报，2017，43(5)：32－41.

5　潘玥斐.深化数字人文研究[N/OL].中国社会科学报，2019－11－25(1)[2021－03－10]. http://news.cssn.cn/zx/bwyc/201911/t20191125_5047664.shtml.

6　邓君，孙绍丹，王阮，等.数字人文视阈下明代科举进士群体时空网络结构分析[J].图书情报工作，2020，64(17)：4－17.

7　史宁昌，李广华，雷勇，等.高光谱成像技术在故宫书画文物保护中的应用[J].文物保护与考古科学，2017，29(3)：23－29.

8　姜育彦，李雅茹.基于数字人文视角的“情感—时空”模型探析[J].农业图书情报学报，2020，32(6)：23－33.

9　李斌，王璐，陈小荷，等.数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例[J].大学图书馆学报，2020，38(5)：72－80，90.

10　Causer T，Grint K，Sichani A M，et al. ‘Making such bargain’：transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription[J]. Digital Scholarship in the Hu-

manities，2018，33(3)：467－487.
11 Abbott A. The ‘time machine’ reconstructing ancient Venice’s social networks[J]. Nature，2017(546)：341－344.
12 陈涛，刘炜，单蓉蓉，等.知识图谱在数字人文中的应用研究[J].中国图书馆学报，2019，45(6)：34－49.
13 Nakamura S. Usage of Japan search RDF model in the development of cultural Japan[J]. デジタルアーカイブ学会誌，2020，4(4)：348－351.
14 Wang X G，Chang W L，Tan X. Representing and linking Dunhuang cultural heritage information resources using knowledge graph[J]. Knowledge Organization，2020，47(7)：604－615.
15 杨海慈，王军.宋代学术师承知识图谱的构建与可视化[J].数据分析与知识发现，2019，3(6)：109－116.
16 陈涛，单蓉蓉，张永娟，等.数字人文研究的语义支撑平台构建研究——以 ECNU-DHRS 平台为例[J].图书馆杂志，2021，40(3)：69－77.
17 王鑫，邹磊，王朝坤，等.知识图谱数据管理研究综述[J].软件学报，2019，30(7)：2139－2174.
18 Berners-Lee T. Linked Data[EB/OL].[2021－02－20].https://www.w3.org/DesignIssues/LinkedData.html.
19 IIIF. For implementers-international image interoperability framework[EB/OL].[2021－02－24]. https://iiif.io/technical-details.
20 刘炜，夏翠娟，张春景.大数据与关联数据：正在到来的数据技术革命[J].现代图书情报技术，2013，(4)：2－9.
21 王昊奋，漆桂林，陈华钧.知识图谱方法、实践与应用[M].北京：电子工业出版社，2019.

作者单位：陈涛，中山大学信息管理学院，广东广州，510006
李惠，南京农业大学人文与社会发展学院，江苏南京，210095
张永娟，中国科学院上海生命科学信息中心，上海，200031
孙安，河南科技大学图书馆，河南洛阳，471000

收稿日期：2021 年 7 月 2 日
修回日期：2021 年 9 月 11 日
（责任编辑：关志英）

LIBRA Technology Theory and Its Application in Historical Image Resources

Chen Tao　Li Hui　Zhang Yongjuan　Sun An

Abstract: Digital Humanities is dedicated to the deep integration of digital technology and humanities research, and the continuous emergence of new technologies is driving the development and transformation of digital humanities. In this paper, we have distilled the five commonly used key technologies from a wide range of digital technologies to form the LIBRA technology system. In LIBRA, the core technologies of Resource Description Framework (RDF) and Linked Data have been prevalent in the field of digital humanities for many years. The International Image Interoperability Framework (IIIF) as an emerging technology has become the mainstream program for image resource research in cultural heritage institutions. Big data and artificial intelligence (AI) technologies have caused great changes in human society, and powerful calculation and intelligent algorithms have given new paradigms and new horizons to human studies. It can be said that LIBRA has become or will become the core technical framework for the construction of digital humanities and will be increasingly valued and implemented. In the case section, the paper explores the in-deep application of LIBRA technology in building multi-dimensional image smart system (MISS) from three aspects: knowledge organization model, data storage model and dynamic text recognition. With the development of technology, and new technologies will complement and reshape the LIBRA system to help the construction and take-off of China’s digital humanities construction.

Keywords: LIBRA; Digital Humanities; International Image Interoperability Framework; Artificial Intelligence; Multi-dimensional Image Smart System