数字人文图像资源语义化建设框架研究*

陈　涛 / 上海图书馆（上海科学技术情报研究所）
张永娟 / 上海大学图书情报档案系，中国科学院上海生命科学信息中心
单蓉蓉 / 上海大学图书情报档案系
刘　炜 / 上海图书馆（上海科学技术情报研究所）

摘　要：图像作为数字人文中非文本资源的主要形式之一，蕴藏着大量宝贵的知识财富。然而长期以来，图像资源由于信息封闭、语义化程度低、资源重复建设且利用率低，形成了信息孤岛、价值洼地。结合IIIF、自然语言处理、OCR、众包、关联数据和知识图谱等技术和概念，构建适合数字人文中图像资源语义化建设的技术框架和模型，提出图像资源数字化重组、概念抽取、语义化注释和智慧化交互的递进实施模式，旨在推动图像资源建设在信息化、数字化和语义化方面的创新转型，对国内图书馆、博物馆等机构的图像资源语义化建设具有参考借鉴价值。

关键词：数字人文　语义化　关联数据　国际图像互操作框架　知识图谱

数字技术与传统人文学科跨界融合的数字人文被视为人文学科繁荣发展的下一个未来，现阶段数字人文研究中非结构化文本数据的处理技术日趋成熟，图像等非文本资源也备受关注。从古代卷轴到现代报纸，从欧洲中世纪手稿到中国书法摹本，这些带有文本的图像蕴藏着大量宝贵的财富，但由于物理对象的衰变、对脚本或语言缺乏了解而造成的阅读困难，资源的交互和深度应用难以实现，形成了严重的信息孤岛。而图像文本资源的语义化组织与建设将丰富数字图书馆元数据建设体系的研究，在数字人文研究领域显得尤为重要。

*本文为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”（19BTQ024）结项成果。

一、相关研究与实践概述

关联数据（Linked Data）是一种简单的语义网实现技术，价值在于通过资源描述框架（Resource Description Framework，以下简称RDF）数据模型，将网络上多数据节点的非结构化数据和采用不同标准的结构化数据转换成遵循统一标准的结构化数据，以便机器理解。数据之间的关联越是丰富，数据的价值就越能得到体现。[①]Christian Bizer指出关联数据是网络中发布和连接不同领域数据的最佳方法，提供了全球化的数据空间。[②]Tom Heath也指出关联数据以分布式、去中心化的思想构建全球化的数据网络（Web of Data）。[③]

关联数据在整合孤立数据、提供开放的元数据服务、实现语义互操作等方面具有广阔的应用前景。在数字人文中，关联数据常用来作为知识组织的方式进行内容增强和基础设施建设，如芬兰数字人文关联开放数据基础设施（LODI4DH）项目旨在创建集中的国家关联数据服务，通过开放接口以结构化、标准化的格式发布和利用数据集，以用于数据密集型数字人文研究，并发布大量的相互关联的核心数据集。曾蕾尝试将智慧数据应用到LAM（图书馆、档案馆、博物馆）等数字人文领域。[④]国内对关联数据的研究也逐渐从理论到落地应用，上海图书馆从2015年开始研究、探索关联数据在数字人文的实践，刘炜、叶鹰等梳理了数字人文的技术和框架体系，[⑤]夏翠娟、林海青等将关联数据应用到古籍循证研究中。[⑥]近几年上海图书馆陆续推出家谱知识库、盛宣怀档案、书目数据发布平台、名人手稿知识库、CBDB关联数据平台[⑦]等一系列数字人文应用平台，并将地名库、人名规范库、机构库等规范数据集作为基础设施开放免费获取。此外，吉林大学

---

①大卫·伍德（David Wood），玛莎·扎伊德曼（Marsha Zaidman），卢克·鲁思（Luke Ruth）、迈克尔·豪森布拉斯（Michael Hausenblas）:《关联数据：万维网上的结构化数据》，蒋楠译，北京：人民邮电出版社，2018年。

②Christian Bizer, Tom Heath and Tim Berners-Lee.,“Linked Data - the Story So Far,” *International Journal of Web & Semantic Technology*, 2009, vol. 5, no. 3, pp. 1-22.

③Tom Heath, Christian Bizer, “Linked Data: Evolving the Web into a Global Data Space,” 1st edition, *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan and Claypool, 2011.

④M. L. Zeng, “Semantic Enrichment for Enhancing LAM Data and Supporting Digital Humanities,” *El profesional de la información*, 2019, vol. 28, no. 1, e280103；曾蕾、王晓光、范炜:《图档博领域的智慧数据及其在数字人文研究中的角色》,《中国图书馆学报》2018年第44卷第1期。

⑤刘炜、叶鹰:《数字人文的技术体系与理论结构探讨》,《中国图书馆学报》2017年第43卷第5期。

⑥夏翠娟、林海青、刘炜:《面向循证实践的中文古籍数据模型研究与设计》,《中国图书馆学报》2017年第43卷第6期。

⑦陈涛、刘炜、单蓉蓉等:《知识图谱在数字人文中的应用研究》,《中国图书馆学报》2019年第45卷第6期。

赵夷平、毕强使用关联数据来进行学术资源网相似文献的发现研究；[①]华东师范大学鲁丹、李欣等也将关联数据和本体应用到数字方志集成平台的构建中。[②]

国际图像互操作框架（International Image Interoperability Framework，以下简称IIIF）旨在通过一组API解决文化资源数字化图像难以被发现、再利用、引用、交换、比较分析等难题，确保全球图像存储的互操作性和可获取性。IIIF的出现为图像资源的知识组织和深度利用提供了新的方法，迅速成为国外GLAM（艺术馆、图书馆、档案馆和博物馆）等机构研究的热点。IIIF基于关联数据的理念，采用JSON-LD进行资源结构的组织，在数字人文中用于馆藏图像资源的展示与交互，如哈佛艺术博物馆的馆藏资源用IIIF进行展示以增强交互性；[③]大英图书馆、盖蒂（Getty）博物馆、欧洲虚拟博物馆（Europeana）、美国艺术档案馆等纷纷使用IIIF来进行馆藏图像资源的展示与分享。IIIF可与W3C的Web注释数据模型结合用于图像资源标注研究，如Gene Loh使用关联数据和IIIF进行图像标注术语词表的管理；[④]利用这些技术，威尔士国家图书馆众包平台进行报纸图片的抄录，大英图书馆LibCrowds项目则进行英国老剧院海报上演员、剧照等信息的标注。国内，上海图书馆率先在家谱、古籍等项目中使用IIIF中的图像API和呈现API来进行大量全文图像资源的展示。

数字人文逐渐使用许多新兴的技术以改变其研究和服务模式，如美国印第安艺术与文化博物馆的“原住民数字档案馆（Indigenous Digital Archive）”项目收集了超过500,000份关于圣达菲印第安学校和其他寄宿学报各种记录和信件的档案文件，该项目在IIIF和Web注释数据模型的基础上，通过光学字符识别（Optical Character Recognition，以下简称OCR）和自然语言处理进行内容的标记来改进访问。瑞士洛桑联邦理工学院（EPFL）数字人文科学实验室研究的威尼斯时光机项目，从水上之城跨越1,000年的地图和手稿出发，利用人工智能、机器学习和虚拟现实技术重塑水城千年历史，成为关联数据应用领域的翘楚。[⑤]

综上所述，关联数据和IIIF这两大国际标准开启了数字人文研究的新时代。国内在关联数据知识库方面的应用已经突破量的积累，正进入质变关键期，但和国外相比，基础设施仍然薄弱，缺乏应用创新性。尤其在图像等非文本资源的语

---

①赵夷平、毕强：《关联数据在学术资源网相似文献发现中的应用研究》，《现代图书情报技术》2016年第3期。
②鲁丹、李欣：《整合异构特藏资源 构建数字人文系统》，《图书馆论坛》2018年第10期。
③Jeffrey P. Emanuel., *Digital Humanities, Libraries, and Partnerships,* Chapter 9, Chandos Publishing, 2018.
④Gene Loh, “Linked Data and IIIF: Integrating Taxonomy Management with Image Annotation,” *2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)* , 2017.
⑤Alison Abbott, “The ‘Time Machine’ Reconstructing Ancient Venice's Social Networks, ” *Nature*, 2017, vol. 546, issue 7658.

义化建设中，需要更为系统化、标准化、自动化的体系结构和框架。

二、图像资源语义化框架设计

根据图像资源的特点，语义化建设框架将分为四步开展：采用IIIF进行图像资源的数字化重组；采用机器学习、众包、实体识别等技术进行图像资源的文本化和概念化抽取；结合开放的关联数据集进行图像资源的语义化注释；结合知识图谱进行图像资源的智慧化交互。

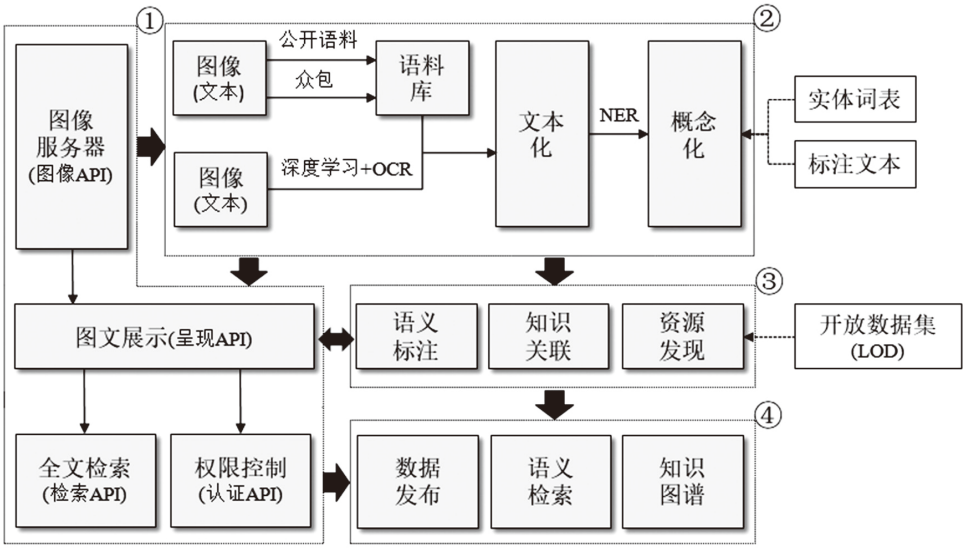

图1　图像资源语义化框架

（一）图像资源的数字化重组

IIIF这一国际通用标准的出现给图像资源的数字化建设提供了新的选择和崭新的用户体验，IIIF提供的四类API可以贯穿图像资源语义化建设的始终。

（1）图像API图像API指定了一套用来获取多种格式图像的请求标准，通过URI来请求图像的区域、大小、旋转、质量特征和格式。图像API可以与OCR技术和众包模式相结合，通过OCR技术将图像资源中的文本识别出来后，通过众包方式让大众参与审核。

（2）呈现API。呈现API（JSON-LD格式）描述了如何以标准方式提供图像的跨域重用，进行复杂的图像结构组织布局。使用呈现API来组织图像资源，可以以跨物理的虚拟组织方式提供全新的资源组织模式，实现在线跨集合（如多

本书中同一实体在不同的图像页，一本书可以看成是不同页的集合）、跨托管机构（如不同机构馆藏的同类图像资源）的图像组织。

（3）检索API。检索API列出了在IIIF中执行文本搜索的互操作机制，但此API本身并不包含语义标注功能，因此要实现图像文本的全文检索，需要和语义标注模型（如OADM：Open Annotation Data Model）结合起来共同使用。

（4）认证API。由于内部策略、法律法规、业务模型等约束，很多馆藏机构要求资源的使用者进行身份验证以获得资源交互的授权。因此可以结合IIIF和LDAP（轻量目录访问协议，Lightweight Directory Access Protocol）控制资源的开放级别和分页。

（二）图像资源的概念抽取

为了实现图像资源语义化建设，需要对图像资源中的实体（概念）进行抽取和识别，并结合实体词表和标注文本共同进行歧义消解、指代消解。

（1）采用众包模式实现语料库收集。图像中的文字识别是语义化建设的基础，开放的和自建的文字语料库，可直接作为图像文本识别的训练集。对于识别困难的古籍，可借助众包思路：将图像文本资源中的文字从图像中切分出来，混杂到较易识别的文字语料中，开放给大众进行校验，如作为某些平台登录时的验证码。达到一定的设置阈值后，即可利用这些图像文本的文字语料，并辅以一定的人工审核来提高文字识别的精度。

（2）基于深度学习的OCR字符识别。使用（1）中收集的语料构建机器学习模型，可以采用卷积神经网络（Convolutional Neural Networks，以下简称CNN）和长短期记忆网络（Long Short-Term Memory，以下简称LSTM）的联合算法，其中CNN用来做图像文本的特征提取，LSTM则用来生成描述。训练时，为了避免多拟合，可以采用多次迭代的方式。训练好模型后，对于图像文本的OCR就转为一个个文字图像的多标签学习问题，可以用来自动识别大量的图像文本资源。

（3）实体标注与抽取。需要借助实体语料库，可以用很多公开的规范词表（人名、地名、时间等）。此外，对于一些无法总结出规范词表的实体，如古地名、建筑名等，可以采用BIO标注，并使用RNN（循环神经网络）-CRF（条件随机场）模型对文本中的实体概念进行抽取，这也是目前深度学习的NER方法中的最主流模型。

（三）图像资源的语义化注释

IIIF框架并不包含语义标注标准和模型，因此要实现图像资源的语义标注，需要结合另一个通用的国际标准“开放标注数据模型（Open Annotation Data Model，以下简称OADM）”，通过注释可以传达有关资源或资源之间关联的信息。OADM规范描述了一种结构化的模型和格式，以使注释能够在不同的硬件和软件平台上共享和重用。语义化注释将实现不同数据集之间资源的发现与知识关联机制，并和IIIF框架进行结合，以知识链接形式关联到外部开放的关联数据集（Linked Open Data，以下简称LOD），从而对之前抽取的实体进行自动化语义标注。语义标注后，抽取出来的每个实体不再以字符串形式存在，而是直接对应到相应的资源URI，而且可以链接到大量的开放数据集。这些开放的数据集组成了实体资源丰富的数据池，对实体内容进行语义层面的增强。

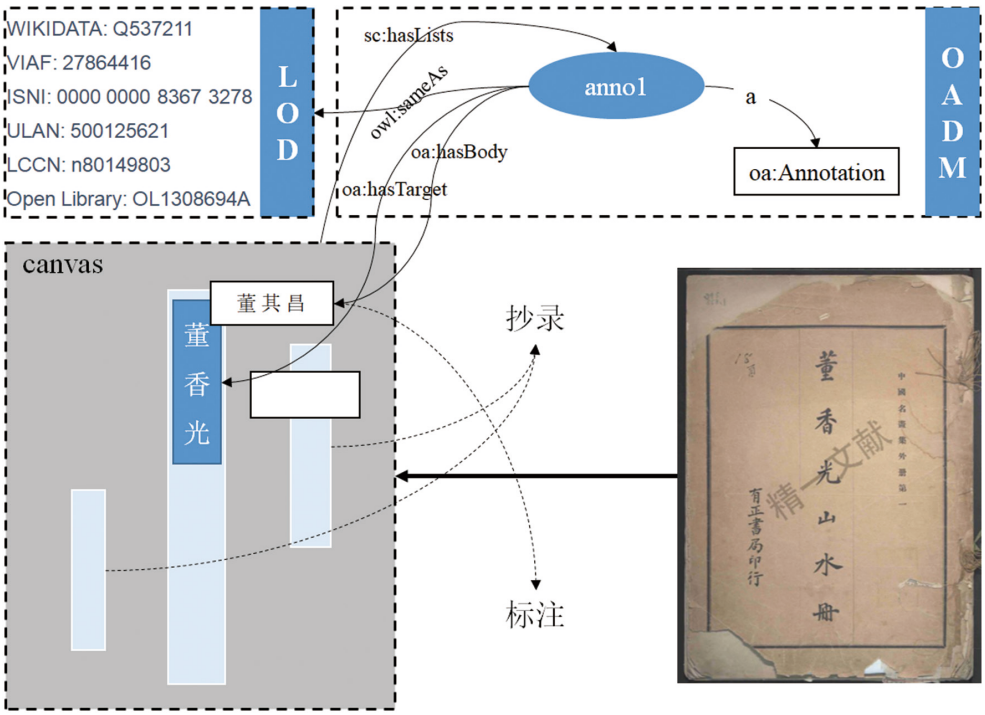

图2　图像资源语义化注释模型

图2显示了IIIF和WADM的集成模型，图像显示在canvas（画布）中，图像中的文字内容称为“抄录”，标注的信息为“标注”，如这里的“董香光”为原图片的文字转录，标注的内容“董其昌”为标注信息。对应到OADM模型中，每一次标注都有其独一无二的ID（URI），通过oa:hasTarget将注释内容链接到图像中的对应区域，同时使用oa:hasBody来指定标注的具体内容，注释本身

也用sc:hasLists属性和IIIF中的canvas进行关联。目前为止的注释，还不能称为语义注释，并没有和外部的数据集进行关联。此时，可以使用关联数据的联邦检索特性，在开放的数据集中检索相关实体资源，如这里的“董其昌”可以在WIKIDATA、VIAF、ISNI、ULAN、LCCN和Open Library数据集中查询到相关的实体链接，有了这些关联链接，就可以在知识图谱中进行知识的融合。

（四）图像资源的智慧化交互

图像资源的数字化建设、概念化抽取和语义化注释将为最终的智慧化交互服务，只有打破图像资源之间的数据鸿沟，结合庞大的开放数据资源才能最大程度释放图像的价值。关于图像资源的应用，有三方面的服务可以提升。

（1）数据发布规范。语义化建设需要秉承分布式、去中心化的数据存储理念，以W3C的开放标准（关联数据发布四原则和开放数据五星模型）为规范发布和共享图像资源及语义标注信息，图像资源才能融入大数据时代的数据潮流，展现大数据时代的知识魅力。

（2）语义检索实现。语义检索是指在知识组织的基础上，从知识库中检索出知识的过程，是一种基于知识组织体系，能够实现知识关联和概念语义检索的智能化的检索方式。图像有了语义注释，就可以通过信息检索、关联数据、语义分析、自然语言处理等技术实现基于语义（概念）的复杂的图像文本信息检索。

（3）知识图谱构建。知识图谱是一种带有语义连接规则的、更规范化的图数据结构，本质上是一种语义网络，常用来揭示实体之间的关系。馆藏机构可以结合自身馆藏资源和外部的开放关联数据集来绘制实体知识图谱（以实体为中心）、集合知识图谱（以藏本为中心）、领域知识图谱（以领域为中心），创新服务模式，提高科研效率与质量。

三、图像语义化关联

图像资源不同于文本资源，图像本身并不能直接用来进行语义化关联，需要通过文本注释，我们在注释时采用了开源的Simple Annotation Server（以下简称SAS）标注服务器。SAS作为标注服务器，集成了IIIF和Mirador图像阅读器，可将注释的信息作为链接数据存储在Apache Jena的RDF数据库中，也可以在Sesame和SOLR中存储注释内容。基于SAS标注服务器，我们做了如下改进：

（1）资源实体语义关联。在进行标注时，当标注的内容为“人”“地”“时间”

等实体时，在将标注内容插入到RDF数据库的同时，对开放数据集中该类实体进行语义关联。

（2）标注内容重建机制。当集成网络中该图像内容的相关标注时，需要将标注内容更新到已有的RDF数据库中，并更新相应的索引数据，以实现对不同来源标注信息的检索。

（3）多资源索引一键生成。SAS可对单一的Manifest文件建立索引，当需要对大量的Manifest文件进行批量索引时，则较为麻烦。实验中改进了索引机制，可以同时批量生成多个Manifest文件索引。

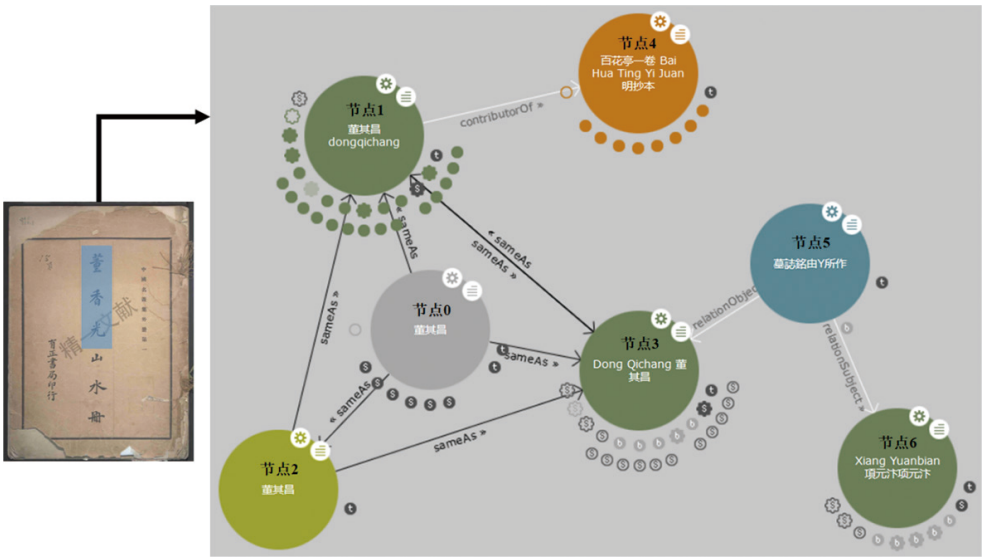

图3　图像资源（关联）知识图谱

图3为图像资源“董香光山水册”首页图像的语义关联示例，通过对“董香光”标注实体内容“董其昌”，可实现与多个开放数据集的内容关联，示例如下：

表1　“董其昌”资源关联节点

数据源	关联资源节点
上海图书馆人名规范库	http://data.library.sh.cn/entity/person/7jfqn5vittx6efhl
中国历代人物传记资料库	http://cbdb.library.sh.cn/entity/person/9dbcvpxlkilf5n3x
维基知识库 WIKIDATA	https://www.wikidata.org/wiki/Q537211
虚拟国际规范文档 VIAF	https://viaf.org/viaf/27864416
GETTY 开放数据集	http://vocab.getty.edu/page/ulan/500125621
美国国会图书馆规范档 LOC	https://id.loc.gov/authorities/n80149803
开放图书馆	https://openlibrary.org/works/OL1308694A
国际标准名称标识	http://www.isni.org/0000000083673278
SinoPedia 百科知识库	http://sinopedia.library.sh.cn/entity/person/34955c1466bd4bb08f293737f9d2483c

有了关联的资源节点后，就可以利用知识图谱方式进行知识融合，图3中节点0为图像标注信息产生的空节点，该节点周边的每个“s”节点对应一个关联的外部数据源，图3中展开了三个国内的开放节点：节点1为上海图书馆人名规范库中的资源节点；节点2为SinoPedia百科知识库中的资源节点；节点3为CBDB中的资源节点，其他的数据节点为国外的数据源。不同的数据源都有自己的特色，比如节点1中含有丰富的文献信息，如古籍、书目信息，节点4为“董其昌”在上海图书馆的古籍信息；节点3中具有大量的人物之间的关系陈述，通过展开节点5，可以看到节点4“董其昌”的墓志铭由节点6“项元汴”所作。

可见，通过对图像资源的语义化建设，可以挖掘大量图像背后的知识，打通图像资源和各种知识库之间的联系。

结论与展望

在研究视角上，我们从资源的基础设施建设着手，瞄准数字图书馆馆藏资源的元数据语义化构建问题，可为图像等非文本资源的语义化建设和开放利用提供新的研究思路和解决方案。在研究内容上，以馆藏中大量的利用率极低的图像文本资源为研究对象，展开图像资源的文本化、概念化以及图像的自动化语义标注与语义应用等方面的研究。在技术方案上，结合关联数据、深度学习、众包、自然语言处理与实体识别、知识图谱等新兴、热门技术，构建了一套图像资源的自动化、语义化生产与应用流程。希望能解决图书馆馆藏图像文本资源信息封闭、语义化程度不高、图像资源重复建设及利用率低等问题，推动图书馆图像资源在信息化、数字化和语义化方面的创新转型。

Research on the Framework of Semantic Construction of Image Resources in Digital Humanities

Chen Tao, Zhang Yongjuan, Shan Rongrong, Liu Wei

Abstract: The images contain a lot of valuable knowledge wealth, which is one of the main forms of non-text resources in digital humanities. However, image resources have long formed a serious information island and low value, because the information is closed, the semantic level is not high, the image resources are repeatedly constructed, and the utilization rate is low. This thesis constructs a technical framework and model suitable for the semantic

construction of digital humanities image resources, combining International Image Interoperability framework (IIIF), natural language processing, OCR, crowd sourcing, Linked Data and Knowledge Graph. The framework proposes a progressive implementation model for digital reorganization, conceptual extraction, semantic annotation, and intelligent interaction of image resources. The purpose is to promote the innovation transformation of image resources in informationization, digitization, semantics, and has practical reference value for the semantic construction of image resources in domestic libraries, museums and other cultural heritage institutions.

Keywords: Digital Humanities; Semanfics; Linked Data; IIIF; Knowledge Graph

（编辑：封帆）