基于本体的简牍数据访问方法研究

王昌龙,高　源,桑秀娟,王玺杰

(西北师范大学 计算机科学与工程学院,甘肃 兰州 730070)

摘　要: 尽管目前已有许多博物馆对所藏的简牍文物开展了数字化工作,但这些工作大多基于传统的关系型数据库和磁盘文件进行数据存储,难以实现知识的互通互连,削弱了数字化文物资源共享的优势。针对该问题,提出一种面向多模态异构简牍数据的集成访问方法。在理论方面,提出面向多模态简牍学的本体构建框架以及基于本体的多模态简牍异构数据集成方案;在实践方面,在本体构建框架的基础上,以"居延新简"为例构建相应的简牍本体。在真实数据上开展实验,从系统灵活性、数据扩展性和单次查询耗时3个方面对集成系统的性能进行评价。结果表明,该系统的平均单次查询耗时约为54 ms,具备一定的实践意义和应用价值。

关键词: 异构;数据集成;本体;简牍;数字人文

DOI: 10. 11907/rjdk. 251054　　扫描二维码阅读全文:

中图分类号: TP392　　**文献标识码:** A　　**文章编号:** 1672-7800(2026)004-0027-08

Research on Ontology-based Data Access Method for Bamboo Slips

WANG Changlong, GAO Yuan, SANG Xiujuan, WANG Xijie

(*College of Computer Science & Engineering*, *Northwest Normal University*, *Lanzhou* 730070, *China*)

Abstract: Although many museums have carried out digital work on their bamboo slips and cultural relics, most of these works are based on traditional relational databases and disk files for data storage, which makes it difficult to achieve knowledge interconnection and weakens the advantages of digital cultural relics resource sharing. A method for integrated access of multimodal heterogeneous bamboo slips data is proposed to address this issue. In terms of theory, propose an ontology construction framework for multimodal bamboo slips and an ontology based heterogeneous data integration scheme for multimodal bamboo slips; In practice, based on the ontology construction framework, take "Juyan New Bamboo Slips" as an example to construct the corresponding bamboo slips ontology. Conduct experiments on actual data to evaluate the performance of the integrated system from three aspects: system flexibility, data scalability, and single query time. The results show that the average query time of the system is about 54 ms, which has certain practical significance and application value.

Key Words: isomerism; data integration; ontology; bamboo slip; digital humanities

0　引言

简牍文物作为古代文明的重要物质载体,具有不可替代的学术价值。对简牍文物实施数字化处理,不仅能够借助先进技术手段实现对其物理状态的精准监测与保护性修护,还能通过构建数字化资源库,为学术研究提供便捷、高效的资料获取途径,从而极大提升简牍文物资源的利用效率与学术研究深度,在文化遗产保护与学术发展领域均具有举足轻重的意义。当前,简牍文物数字化工作大多基于传统的关系型数据库和磁盘文件进行数据存储。这种方式在数据结构与关联性处理上存在明显局限,难以实现知识的互通互连,削弱了数字化文物资源共享的优势。为解决这一问题,利用知识关联实现异构简牍数据集成,进而提供一种更有效的访问方法已经成为简牍数字化领域的重要研究任务。

收稿日期: 2025-01-21

基金项目: 国家自然科学基金项目(62362060,72364033)

作者简介: 王昌龙(1977-),男,博士,西北师范大学计算机科学与工程学院副教授、硕士生导师,研究方向为知识表示与推理、知识图谱、智能软件工程;高源(1994-),男,西北师范大学计算机科学与工程学院硕士研究生,研究方向为智能软件理论与技术;桑秀娟(2000-),女,西北师范大学计算机科学与工程学院硕士研究生,研究方向为智能计算与软件;王玺杰(1997-),男,西北师范大学计算机科学与工程学院硕士研究生,研究方向为智能软件理论与方法。本文通讯作者:高源。

1 相关研究

在利用知识关联性对文物数字化数据进行集成的研究中，李贺等[1]提出基于命名实体识别模型与实体关系抽取模型构建简帛医药知识的图数据，最终实现了简帛医药知识检索与可视化。谢玮等[2]提出一种面向《天工开物》的数字资源集成方法，通过构建本体和关联数据的方法搭建《天工开物》图像及版本知识图谱，从而得到一个具备多重路径知识检索功能的共享知识库。熊晶等[3]提出一种基于多源异构甲骨数据的集成方法，该方法融合基于文献计量学的科学知识图谱与基于知识库的知识图谱，构建出一个包含148 305个实体和434 032条关系的甲骨学知识图谱。周冬艳等[4]针对青瓷文物的特征及相关知识进行分类研究，借助自然语言处理工具构建了青瓷知识图谱。胡汗林等[5]提出一种对青铜器概念与术语进行分析，并利用Neo4j数据库构建青铜器知识图谱的方法。梁杨等[6]以泸县宋代石刻纹样为载体，对宋代石刻纹样知识体系进行挖掘，并基于此构建了宋代石刻纹样知识图谱。Stefano等[7]以知识图谱的方式对意大利文化遗产数据进行集成。Daphne等[8]提出一种基于本体对文物及其相关数字资源进行关联并构建知识图谱的方法。Ikrom等[9]将地理空间语义网与文化遗产数据相融合，构建出利用机器可读和可处理的资源描述框架所描述的文化遗产地理空间语义网。此外，台湾史语所基于所藏的13 000枚居延汉简开发了“简牍字典”在线检索系统，该系统构建了简牍学本体、汉字本体和标注本体，并在此基础上提供了诸如释文内容的关键字标记、简牍集的虚拟组织编排、图像查看与对比等功能，以便帮助学者完成文字释读及简册复原等工作任务[10]。上述成果虽然实现了以知识对文物数据进行关联的目的，但均需要获得统一的数据类型，无法直接对已有异构数据进行集成，因此不可避免地产生了物化成本，不利于系统扩展。

基于本体的数据访问（Ontology-Based Data Access，OBDA）技术也称为虚拟知识图谱（Virtual Knowledge Graph）技术[11]。该技术能够在不改变原有数据存储管理的情况下使用本体提供更抽象的数据访问视图。其突出优点为能够充分利用数据库和文件管理系统中成熟高效的数据组织和索引功能，具备较好的可扩展性。例如，Calvanese等[12]提出的EPNet利用OBDA技术构建了面向古罗马帝国商业和贸易文物数据的虚拟知识图谱系统。受到该研究的启发，本文结合OBDA技术与多模态知识图谱（Multi-Modal Knowledge Graph，MMKG）技术[13]，基于“居延新简”数字化工作所产生的数据提出一种面向异构简牍数据的集成方法，旨在解决传统数据集成方法对于文本、图像等异构简牍数据集成不完整，难以使普通用户直观便捷地访问异构数据源的问题。

2 异构简牍数据特点分析

2.1 关系型简牍数据特点分析

在简牍数字化工作中，关系型数据通常是采集数量最多的一部分。根据简牍文物的特征，这些数据通常包含简牍的编号、通称、朝代和形制等信息。以“居延新简”的数字化工作为例，其中编号为“EPT1:1”的简采集到的部分基本信息如表1所示。该简牍集合的共性信息均按照关系型数据模型存储于名为“BAMBOO_BISIC”的表内，其中的某条数据如表2所示。

Table 1 Part of the basic information contained in the document numbered EPT1:1 in the Ju Yan Xin Jian

表1 “居延新简”中编号为“EPT1:1”的简所包含的部分基本信息

编号	通称	朝代	年代（公元）	形制	集
EPT1:1	居延新简	汉代	-128—32	简	破城子探方一

Table 2 A certain piece of data stored in a relational database named "BAMBOO_BISIC" table

表2 存储于关系型数据库中名为“BAMBOO_BISIC”表内的某条数据

字段名	数据类型	值
ID	varchar(36)	1cbfbee7-fffe-4f91-9ed9-80dc8bf6ce6e
NAME	varchar(20)	居延新简——甲渠候官
TIME_AD	varchar(20)	-128—32
TIME_CC	varchar(20)	汉代
SAVE_LOCATION	varchar(40)	甘肃简牍博物馆

2.2 文本型简牍数据特点分析

简牍的数字化工作通常会产生许多非结构化文本数据。“居延新简”的数字化数据包括“释文”、“校释”和“集释”等。以编号为“EPT1:1”的简为例，其正面释文为：“令使……令使三人尉使四人名如牒请所用代衰主官”，这段释文对应的校释为“「衰」，文物本作「褒」，此从中华本。按，「衰」「褒」异体字，简文中多用于人名。此字，文物本皆释作「褒」，后文悉红外线图版释定，不俱出注。第二个「令」字左部磨灭无存，系整理者拟释，可从”。此外，专家还对这段释文中的“令使”“尉使”和“牒”3个词分别进行了集解。在简牍的数字化工作中，上述数据通常会以文本文件的格式存储于分布式计算机系统的硬盘介质中。

2.3 影像型简牍数据特点分析

简牍的数字化工作还会采集大量影像信息，如每枚简牍的彩色影像和红外影像。这些影像信息均按照影像类别及其采集部位进行分类和标注，并以图像文件的格式存储于计算机系统的硬盘介质中。例如，编号为EPT1:1（正面）、EPT2:1（正面）和EPT2:1（背面）的简的彩色图像如图1所示。

3 简牍学本体构建

为提高数据的可重用性和可交换性，在为特定领域构建本体时应采用统一的模型。目前，文化遗产领域已有一

Fig. 1　Color images of the bamboo slips numbered EPT1:1 (front), EPT2:1 (front) and EPT2:1 (back)

图 1　编号为 EPT1:1(正面)、EPT2:1(正面)和 EPT2:1(背面)的简的彩色图像

些较为成熟的本体模型，包括 CIDOC CRM[14]、ABC Ontology[15]和 RKD[16]等。其中，CIDOC CRM 是严格遵循本体原则的一个模型，至今已拥有 20 年的开发与维护历史，其最新官方版本(ISO Correspondence)为 7.1.3。国内外已有很多研究成功应用 CIDOC CRM 模型构建了文化遗产领域本体，例如文献[10]和文献[17-20]。该模型能将各种碎片知识相连接，具备较为明显的优势。因此，本文利用 CIDOC CRM 模型的可扩展性，使简牍数字化数据成为结构化的语义知识单元，进而完成简牍学本体的构建。

在本体构建过程中，常用方法包括 IDEF5 法、骨架法、TOVE 法、METHONTOLOGY 法、KACTUS 法、SENSUS 法、七步法和循环获取法等[21]。本文采用关键概念自动抽取与领域专家构建相结合的方式对 CIDOC CRM 模型进行精简与扩展，并自下向上地构建多模态简牍学本体模型，基于 KACTUS 法尝试性提出一种多模态简牍学本体构建框架，优势在于注重对现有知识进行重用，并且同时适用于自动构建和手工构建，主要应用于网络领域。该框架可分为 5 个阶段，具体如图 2 所示。

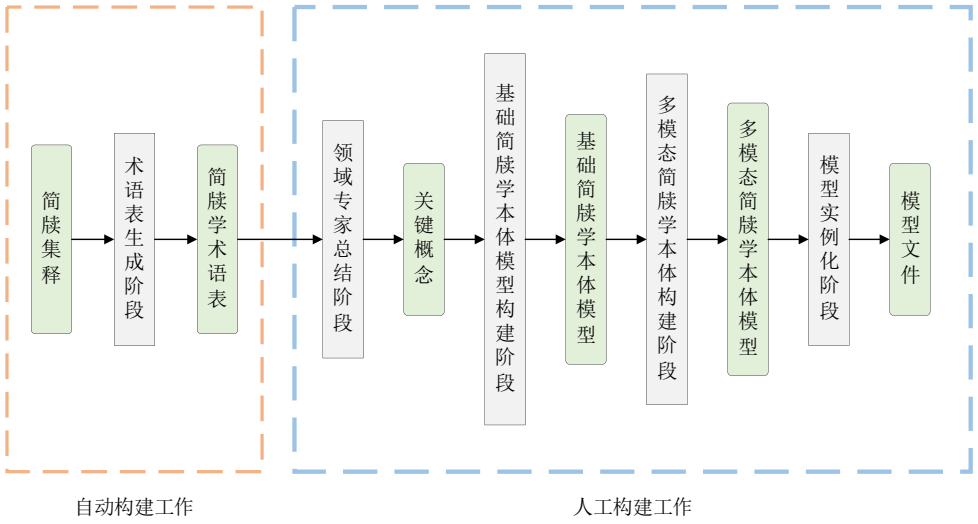

Fig. 2　Construction process of multi-modal knowledge ontology model for bamboo slips

图 2　多模态简牍学本体构建框架

3.1　术语表生成阶段

该阶段选取目标简牍的相关集释作为概念抽取源，利用 HanLP 工具对其中的文本内容进行关键词短语提取，从而得到带有语义权重的初始语料表[22]。为确保简牍学关键概念最大覆盖的临界语义权重，首先选取 10% 作为语义权重临界值的增量，分别过滤出初始语料表中语义权重不低于 10% 至不低于 90% 的关键词短语，构成 9 个精简关键词短语子集，外加初始语料表本身(可视为由语义权重不低于 0% 的关键词短语构成)作为对照用关键词短语子集。接着，计算这 10 个子集的关键概念覆盖率 $\mathcal{R}_c$。公式为：

$$\mathcal{R}_c = \frac{\Sigma(c')}{\Sigma(c)} \times 100\% \tag{1}$$

式中：c 为某一精简关键词短语子集；c' 为由领域专家随机提供的一组由 10 个简牍学中常见关键概念构成的参照名词子集 $\mathcal{C}$ 与集合 c 的交集；$\Sigma(c)$ 为集合 c 中包含的关键词短语总数；$\Sigma(c')$ 为集合 c' 中包含的名词数量。$\mathcal{R}_c$ 值越大，表示该精简关键词短语子集对简牍学关键概念囊括得越全面。

为在保证简牍学关键概念最大程度覆盖的前提下尽可能排除不属于简牍学关键概念的关键词短语，本文还计算了这 10 个精简关键词短语子集的关键概念浓度 $\mathcal{C}_c$。公式为：

$$\mathcal{C}_c = \frac{\Sigma(r')}{\Sigma(r)} \times 100\% \tag{2}$$

式中：r 为从某精简关键词短语子集 c 中随机抽取出的数量为集合 c 中 10% 元素的关键词短语集合；r' 为领域专家于 r 中确定的与简牍学关键概念相关的关键词短语组成的子集；$\Sigma(r)$ 为集合 r 中包含的关键词短语数量；$\Sigma(r')$ 为集合 r' 中包含的关键词短语数量。$\mathcal{C}_c$ 值越大，表示该精简关键词短语子集中与简牍学关键概念无关的关键词短语越少。

为确保构建出的本体模型能够在尽可能精简的前提

下完整描述简牍学领域的关键概念，在确定语义权重临界值时，应选取能够令所得精简关键词短语子集的$\mathcal{R}_c$值等于100%且$\mathcal{C}_c$值尽可能高的语义权重作为临界值。以“居延新简”的相关集释《居延新简集释》为例，本文对上述10个关键词短语子集分别进行3次独立实验，计算出每个语义权重临界值对应的关键词短语子集的$\mathcal{R}_c$值和$\mathcal{C}_c$值。实验结果如图3和图4所示。

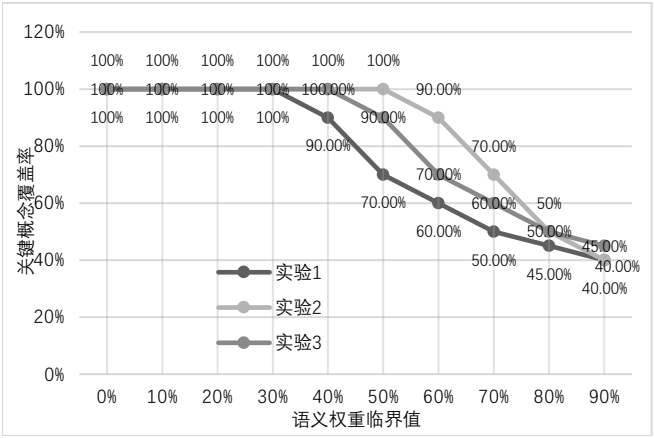

Fig. 3 The $\mathcal{R}_c$ value of subsets of keyword phrases filtered using different semantic weight threshold values

图3 使用不同语义权重临界值过滤出的关键词短语子集的$\mathcal{R}_c$值

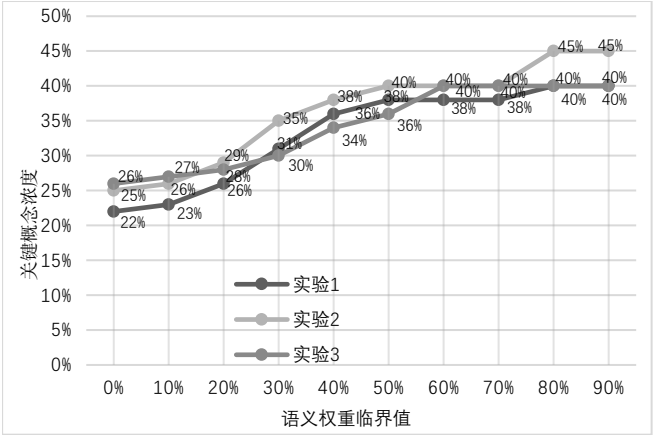

Fig. 4 The $\mathcal{C}_c$ value of subsets of keyword phrases filtered using different semantic weight threshold values

图4 使用不同语义权重临界值过滤出的关键词短语子集的$\mathcal{C}_c$值

根据上述结果，当使用30%作为语义权重临界值时，所得精简关键词短语子集能在完全囊括简牍学中常见关键概念的前提下使与简牍学关键概念无关的关键词短语尽可能少。据此，本文选取30%作为语义权重临界值，从初始语料表中过滤出语义权重大于等于该值的关键词短语，从而获得所含关键词短语比初始语料表更为精简的简牍学术语表。

3.2 领域专家总结阶段

在简牍学术语表作为参考的基础上，结合专家意见归纳出简牍学中涉及的关键概念。

3.3 基础简牍学本体模型构建阶段

基于总结出的关键概念对CIDOC CRM模型进行概念对应与扩充，从而得到基础简牍学本体模型。

3.4 多模态简牍学本体构建阶段

为使该本体模型能够表示文本、图片等多模态信息，利用MMKG技术对其进行进一步扩充。在传统的知识图谱中，数据被定义为有向图。表示为：

$$\mathcal{G}=\{\mathcal{E},\ \mathcal{R},\ \mathcal{A},\ \mathcal{V},\ \mathcal{T}_{\mathcal{R}},\ \mathcal{T}_{\mathcal{A}}\} \tag{3}$$

式中：$\mathcal{E}$、$\mathcal{R}$、$\mathcal{A}$、$\mathcal{V}$分别为实体、关系、属性和属性值的集合；$\mathcal{T}_{\mathcal{R}}=\mathcal{E}\times\mathcal{R}\times\mathcal{E}$和$\mathcal{T}_{\mathcal{A}}=\mathcal{E}\times\mathcal{A}\times\mathcal{V}$分别为关系三元组和属性三元组的集合；三元组$(s,p,o)\in\mathcal{T}_{\mathcal{R}}$用于表示实体$s\in\mathcal{E}$与实体$o\in\mathcal{E}$之间的关系，$p\in\mathcal{R}$。

MMKG技术将多模态信息引入知识图谱，即在有向图$\mathcal{G}=\{\mathcal{E},\ \mathcal{R},\ \mathcal{A},\ \mathcal{V},\ \mathcal{T}_{\mathcal{R}},\ \mathcal{T}_{\mathcal{A}}\}$中加入多模态信息。MMKG拥有两种表示模式：一种是将多模态数据作为实体或概念的特定属性值，该种方式表示的MMKG被称为A-MMKG。其定义为：

$$\mathcal{G}=\{\mathcal{E},\ \mathcal{R},\ \mathcal{A},\ \mathcal{V},\ \mathcal{T}_{\mathcal{R}},\ \mathcal{T}_{\mathcal{A}}'\} \tag{4}$$

$$\mathcal{T}_{\mathcal{A}}'=\mathcal{E}\times\mathcal{A}\times\left(\mathcal{V}_{\mathcal{KG}}\cup\mathcal{V}_{\mathcal{MM}}\right) \tag{5}$$

式中：$\mathcal{V}_{\mathcal{KG}}$为知识图谱属性值的集合；$\mathcal{V}_{\mathcal{MM}}$为多模态数据的集合。在A-MMKG中，由于多模态数据被视为属性值，在三元组(s,p,o)中，s表示实体，o表示其对应的多模态数据之一。例如，当o为图像信息时，关系p则为“hasImage”。

另一种表示模式是将多模态数据作为知识图谱中的实体，该种方式表示的MMKG被称为N-MMKG。其定义为：

$$\mathcal{G}=\{\mathcal{E},\ \mathcal{R},\ \mathcal{A},\ \mathcal{V},\ \mathcal{T}_{\mathcal{R}}',\ \mathcal{T}_{\mathcal{A}}\} \tag{6}$$

$$\mathcal{T}_{\mathcal{R}}'=\left(\mathcal{E}_{\mathcal{KG}}\cup\mathcal{E}_{\mathcal{MM}}\right)\times\mathcal{R}\times\left(\mathcal{E}_{\mathcal{KG}}\cup\mathcal{E}_{\mathcal{MM}}\right) \tag{7}$$

式中：$\mathcal{E}_{\mathcal{KG}}$为知识图谱实体的集合；$\mathcal{E}_{\mathcal{MM}}$为多模态数据的集合。由于多模态数据被视为新的实体，这种表示模式可以表示更多的模态间和模态内关系。

A-MMKG模式虽然在表示多模态信息上更加简洁，但是无法有效表示多模态信息本身附带的相关属性信息。由于简牍的多模态信息通常包含与该信息相关的附加属性信息，如影像信息的长度、宽度、类型等，本文选择N-MMKG模式来表示简牍学本体中的多模态信息。与A-MMKG模式相比，N-MMKG模式不仅能够表示多模态信息本身，而且能灵活表示其附带的相关属性信息。

3.5 模型实例化阶段

使用Protégé工具对本体模型进行实例化，并使用RDF格式进行存储。其中，学科通用实体类均从CIDOC CRM本体族中复用，而涉及简牍学具体概念和专业描述的实体类则在CIDOC CRM模型的基础上进行自定义，整体依旧遵循CIDOC CRM本体模型的层次逻辑。本文以“居延新简”为例，利用上述框架构建出一个简牍学本体模型。该模型的实体类层次图如图5所示。

此外，通过对简牍学关键概念进行梳理与语义对应，

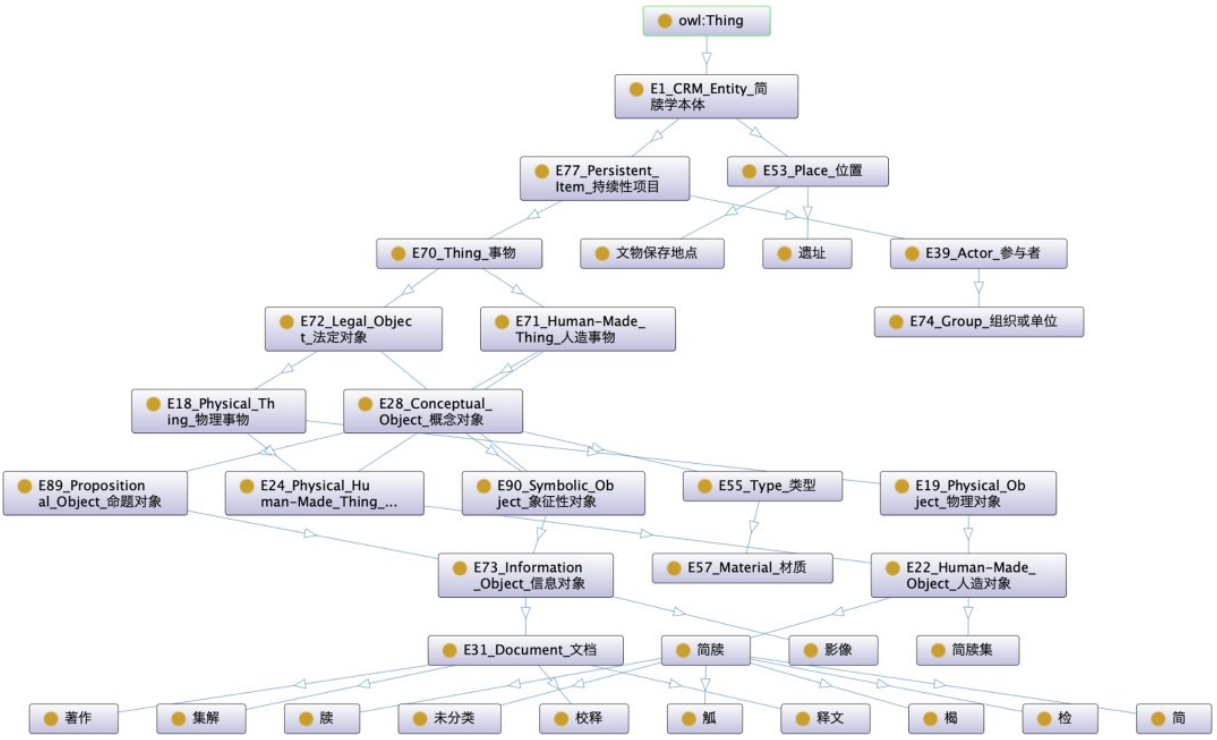

Fig. 5　Entity class hierarchy diagram of ontology model for bamboo slips

图5　简牍学本体模型的实体类层次图

在该模型中定义了12对(24种)对象属性和37种数据属性。部分内容如表3和表4所示。在传统的关系型数据库中,两个字段之间没有知识关联。而本文构建的简牍学本体模型能表示数据之间的关系。例如,“简牍”与“E57_Material_材质”之间由关系“材质为”连接,这样表示的概念具有上下文,赋予了其现实含义。相较于传统的本体构建流程,本文框架的特点在于将自动化构建与手工构建相结合,在降低构建成本的同时确保了本体模型的完善性。

Table 3　Partial object attributes in the ontology model of bamboo slips

表3　简牍学本体模型中的部分对象属性

Name	Domains	Ranges	Disjoint with
包含	简牍集	简牍	属于
属于	简牍	简牍集	包含
关联影像	简牍	影像	关联简牍
关联简牍	影像	简牍	关联影像

Table 4　Partial data attributes in the ontology model of bamboo slips

表4　简牍学本体模型中的部分数据属性

Name	Domains	Ranges
通称	简牍集	xsd:string
影像宽度	影像	xsd:int
重量_g	简牍	xsd:float
信息对象URI	E73_Information_Object_信息对象	xsd:anyURI

4　基于简牍本体的数据集成

在OBDA技术中,数据集成是通过声明本体与数据源之间的映射来实现的。其中,映射由多组断言组成,每组断言均与本体的概念或属性相关联,即对应数据源的一个结构化查询语言(Structured Query Language,SQL)查询。当用户在OBDA系统中执行一次SPARQL查询时,首先,系统会根据知识本体的实体类定义对SPARQL查询的内容进行重写;其次,根据映射将重写后的查询展开为对数据源的同义SQL查询,从而获得SQL查询结果;最后,系统会再次根据映射对查询结果进行翻译,从而将基于图结构的SPARQL查询结果返回给用户。OBDA技术并不会构建出一个真正的知识图谱,而是在不改变数据源内容和存储状态的情况下对其进行集成,并将用户对于本体的SPARQL查询转换为对数据源的SQL查询。相较于构建出一个实际的领域知识图谱,该方法不仅能够有效减少物化成本,而且能充分发挥关系型数据库在高查询效率、强安全性和稳健的事务支持等方面的优势[11]。

目前已有的OBDA系统包括D2RQ[23]、Ontop[24]和Oracle Big Data Spatial and Graph[25]等。其中,Ontop系统由博赞博尔扎诺自由大学开发,并得到了Ontopic公司的商业支持。该系统使用Apache 2许可证授权,支持R2RML、SPARQL 1.0,并且针对JOIN、UNION以及OPTIONAL运算符进行了优化。Ontop系统不仅实现了基于OWL2QL的本体推理功能,而且提供了处理空间数据、时间数据和非关系型数据(例如JSON文档)的原型扩展。本文利用Ontop系统作为中间件设计了基于本体的异构简牍数据集成框架。其结构如图6所示。

该框架主要由以下4个部分组成:①异构数据源。简牍数字化工作产生的基本信息数据库和各类相关文件分

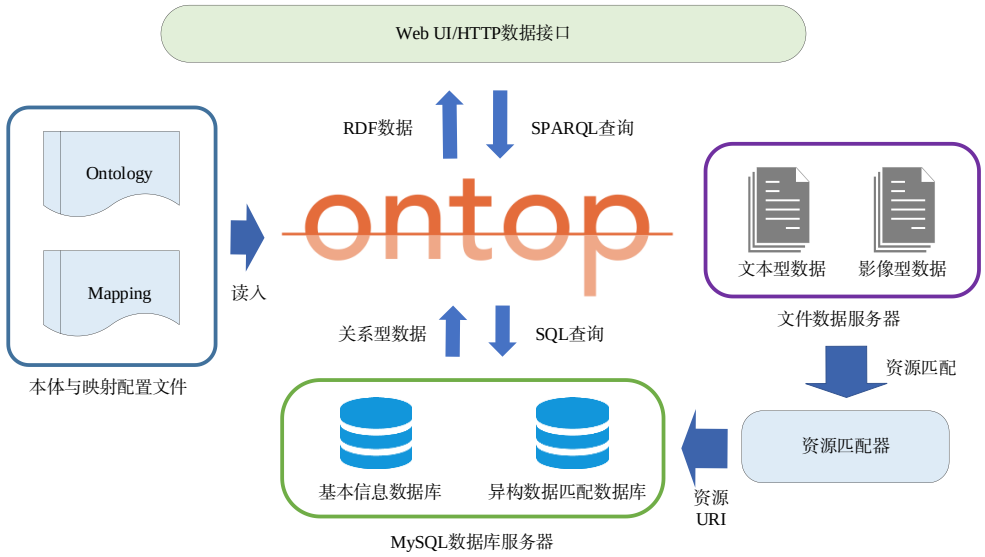

Fig. 6 Ontology-based heterogeneous bamboo slip data integration framework

图 6 基于本体的异构简牍数据集成框架

别存储于独立运行的 MySQL 数据库服务器和支持 WebDAV 协议的文件数据服务器上，作为该系统的异构数据源；②资源匹配器。该部分负责按照预先设定的匹配规则将文件数据与简牍基本信息进行关联，并将文件服务器为该文件生成的 URI 与其对应的简牍基本信息存入多源异构数据匹配数据库的相应表中，以便进行映射。资源匹配器遵循特定的接口，能根据实际需求在系统中进行替换，进而增强系统的扩展性；③中间层。该部分通过将构建好的本体文件及其对应的数据源映射规则文件输入以终端模式部署的 Ontop 系统中，实现对所有异构数据的集成，其中映射文件可根据不同数据源的结构进行相应修改，进一步增强系统的扩展性；④对外接口层。该部分负责对外提供服务，在遵循 RESTful 风格的基础上，通过 http 协议向用户分别提供 Web UI 操作和应用程序调用服务的接口。

本文以“居延新简”为例，利用上述集成框架搭建出一个面向该简牍数字化数据的集成系统。如图 7 所示，当用户在该系统的 Web UI 上查询基本信息时，系统会以基本数据类型直接返回结果内容；如图 8 所示，当查询结果为文件时，系统会以目标文件对应的 URI 返回结果内容，用户可以进一步根据该 URI 访问存储于分布式文件数据服务器上的资源。该集成框架的主要优势为在不改变原有系统数据存储和管理模式的前提下，实现利用知识关联性进行集成的目的，同时还能充分发挥原有系统的稳定性和扩展性。

5 性能评价

本文采用 EPNet 项目进行性能评价时所采用的评价标准，分别从系统灵活性、性能扩展性和单次查询耗时 3 个方面进行评价[14]。

Fig. 7 Using SPARQL to query the material of EPT1:1

图 7 使用 SPARQL 查询编号为 EPT1:1 的简的材质

5.1 系统灵活性

从软件系统的角度看，本文系统中的数据源模块、配置模块、集成模块以及接口模块均具有较强的独立性，而各模块之间存在较松的耦合关系，数据的存储管理对于终端用户而言完全透明，因此灵活性较高。

5.2 数据扩展性

在 OBDA 框架中，本体的主要作用是为用户提供一致的访问视图。而用户查询则通过映射转换，最终提交由各数据源系统的查询引擎进行处理。现有数据系统（如关系型数据库和分布式文件系统）具备技术成熟性优势，因此本文方法亦具有较好的数据扩展性。

Fig. 8　Using SPARQL to query the color image of the back side of EPT2:1

图8　使用SPARQL查询编号为EPT2:1的简的背面彩色影像

5.3　单次查询耗时

为对单次查询耗时进行评价,本文设计了4个复杂度由低到高的SPARQL查询,分别为:

查询1　查询所有类型为简的简牍编号

```
SELECT ? 简 ? 编号 {
? 简 a :简 ; :编号 ? 编号 .
}
```

查询2　EPT1:1的材质是什么

```
SELECT ? 材质名称 WHERE {
? 简牍 :编号 'EPT1:1' .
? 简牍 :材质为 ? 材质 .
? 材质 :材质名称 ? 材质名称 .
}
```

查询3　EPT1:1对哪些原文做了集解

```
SELECT ? 所在面 ? 原文 ? URI WHERE {
? 简牍 :编号 'EPT1:1' .
? 简牍 :集解为 ? 集解 .
? 集解 :集解所在面 ? 所在面 .
? 集解 :集解原文 ? 原文 .
? 集解 :信息对象URI ? URI .
}
```

查询4　“居延新简”中包含多少条简牍信息

```
SELECT (COUNT(? 简牍) AS ? 总数) WHERE
{
? 简牍集 :通称 '居延新简' .
? 简牍集 :包含 ? 简牍 .
}
```

分别在前文系统的Web UI和原始数据源的SQL客户端中执行上述SPARQL查询和同义的SQL查询,最终得到各个查询的耗时如表5所示。可以看出,在集成系统中执行SPARQL查询的耗时均未超过1 s,平均耗时为54 ms,说明使用该方法进行数据集成后的单次查询耗时能够满足实际应用需求。而在与直接执行同义SQL查询的耗时对比中,所有测试查询的耗时均有所增加。其中,最小增加值出现在查询编号1的测试中,增加值为30 ms;最大增加值出现在查询编号为4的测试中,增加值为55 ms。分析其原因,耗时增加的程度主要与查询条件的复杂度相关,OBDA技术在原数据源执行查询流程的基础上额外增加了重写、展开、翻译等操作。在实际应用中,可以认为使用该方法进行数据集成后的查询耗时均比未集成前有所增加。

Table 5　Time consumption of each query

表5　各查询耗时　　**(ms)**

查询编号	SPARQL耗时	SQL耗时
1	42	12
2	53	13
3	55	16
4	69	14

6　结语

本文首先基于CIDOC CRM模型构建出一个多模态简牍学本体,然后利用Ontop系统构建了一个能集成异构简牍数据的系统,并从系统灵活性、数据扩展性和单次查询耗3个方面对其性能进行了评价。该系统可在不改变源数据结构和存储方式的前提下将这些数据以知识图谱的形式集成起来,并对外提供访问接口,使用户以一致的方式进行搜索与访问。未来计划从以下3个方面继续开展工作:①系统的实践应用。将该系统实际应用于“居延新简”的在线检索系统中,并根据其在实际应用中的表现和用户满意度进行完善;②数据丰富工作。借助网络爬虫和数据抽取技术对互联网上简牍相关的非结构化数据进行集成,扩大数据访问范围;③数据可视化访问。在该方法的基础上研究简牍知识可视化查询访问,以便用户能够更为直观地访问系统中的图像、文本甚至网页等异构数据。

参考文献:

[1] LI H, ZHU L L, LIU J Y. Research on ontology based knowledge organization of bamboo and silk medicine[J]. Library and Information Service, 2022,66(22): 16-27.
李贺, 祝琳琳, 刘嘉宇. 基于本体的简帛医药知识组织研究[J]. 图书情报工作, 2022, 66(22): 16-27.

[2] XIE W, HENG Y, QIU J X. Research on the application of knowledge graph for image resources in the "Tiangong Kaiwu" version[J]. Packaging Engineering, 2023, 4(S1): 480-535.
谢玮, 衡雨, 邱菊芯. 面向《天工开物》版本图像资源的知识图谱应用研究[J]. 包装工程, 2023, 4(S1): 480-535.

[3] XIONG J, JIAO Q J, LIU Y T. Oracle bone studies knowledge graph construction based on multi-source heterogeneous data[J]. Journal of Zhejiang University(Science Edition), 2020, 47(2): 131-150.
熊晶, 焦清局, 刘运通. 基于多源异构数据的甲骨学知识图谱构建方法研究[J]. 浙江大学学报, 2020, 47(2): 131-150.

[4] ZHOU D Y, GAO L F, WU Y F. Research on celadon knowledge map construction based on ontology [J]. Library Journal, 2024, 43 (3): 91-100.
周冬艳, 高鲁放, 吴艳芳. 基于本体的青瓷知识图谱构建方法研究[J]. 图书馆杂志, 2024, 43(3): 91-100.

[5] HU H L, DENG S H. Application of Knowledge graph in the construction of bronze digital collection[J]. Digital Library Forum, 2023, 19(4): 1-8.
胡汗林, 邓三鸿. 知识图谱在青铜器数字馆藏建设中的应用[J]. 数字图书馆论坛, 2023, 19(4): 1-8.

[6] LIANG Y, CUI X, ZHANG C. Research on knowledge system mining and knowledge map construction of stone carving patterns in Song Dynasty[J]. Journal of Communication University of China: Science and Technology, 2022, 29(4): 41-49.
梁杨, 崔鑫, 张骋. 宋代石刻纹样知识体系挖掘与知识图谱构建研究[J]. 中国传媒大学学报(自然科学版), 2022, 29(4): 41-49.

[7] STEFANO F, ANDREA L, PAOLA V. A large interlinked knowledge graph of the Italian cultural heritage[C]//Proceedings of the 13th Conference on Language Resources and Evaluation, 2022: 6280-6289.

[8] DAPHNE K M, EVGENIA V. The use of ontologies for creating semantic links between cultural artifacts and their digital resources[C]//10th International Symposium on the Conservation of Monuments in the Mediterranean Basin, 2018: 541-545.

[9] IKROM N, ERIK C, DAVID A M. A survey of geospatial semantic web for cultural heritage[J]. Heritage, 2019, 2(2): 1471-1498.

[10] CHEN S J. Design of interoperability in digital humanities: a case study of the interpretation and restoration of the Han Dynasty wooden slips from Edsen-Gol[J]. Journal of Educational Media & Library Sciences, 2021, 58(2): 193-235.

[11] GUO H X, LIN F D, BENJAMIN C. Virtual knowledge graphs: an overview of systems and use cases[J]. Data Intelligence, 2019, 1(3): 201-223.

[12] CALVANESE D, LIUZZO P, MOSCA A. Ontology-based data integration in EPNet: production and distribution of food during the Roman Empire[J]. Engineering Applications of Artificial Intelligence, 2016, 51: 212-229.

[13] PENG J H, HU X Y, HUANG W B. What is a multi-modal knowledge graph: a survey[J]. Big Data Research, 2023, 32: 100380.

[14] ICOM/CIDOC. Volume A: definition of the CIDOC conceptual reference model[EB/OL]. https://cidoc-crm.org/sites/default/files/Documents/cidoc_crm_version_7.1.3.html.

[15] CARL L, JANE H. The ABC ontology and model[C]//Proceedings of the International Conference on Dublin Core and Metadata Applications, 2001: 160-176.

[16] Netherlands Institute for Art History. Home | RKD-Netherlands Institute for Art History[EB/OL]. https://www.rkd.nl/en.

[17] LIU H Z, BAO H, YU J H. A semantic web architecture of virtual museum based on CIDOC CRM[J]. Application Research of Computers, 2006(4): 50-53.
刘宏哲, 鲍泓, 余杰华. 基于CIDOC CRM的虚拟博物馆语义网络架构[J]. 计算机应用研究, 2006(4): 50-53.

[18] DAI T. Research on metadata specification of digital image of museum cultural relics based on CIDOC CRM—taking the design of cultural relics image metadata system of National Museum of China as an example[J]. Chinese Museum, 2020(3): 131-136.
戴畋. 基于CIDOC CRM的博物馆文物数字化影像元数据规范研究——以中国国家博物馆文物影像元数据体系设计为例[J]. 中国国家博物馆, 2020(3): 131-136.

[19] DIMITRIS K, GEORGIOS G. Incorporate cultural artifacts conservation documentation to information exchange standards-the DOC-CULTURE case [J]. Procedia-Social and Behavioral Sciences, 2014, 147: 495-504.

[20] DAVIDE V, DORA M, IRENE P R. A tool to explore the population of a CIDOC-CRM ontology [J]. Procedia Computer Science, 2021, 192: 158-167.

[21] LI A H, XU Y Z, CHI Y X. Review of ontology construction and applications[J]. Information Studies: Theory & Application, 2023, 46(11): 189-195.
李爱华, 徐以则, 迟钰雪. 本体构建及应用综述[J]. 情报理论与实践, 2023, 46(11): 189-195.

[22] Natural Semantics (Qingdao) Technology Co., Ltd. HanLP[EB/OL]. https://www.hanlp.com/semantics/dashboard/index.
自然语义(青岛)科技有限公司. HanLP[EB/OL]. https://www.hanlp.com/semantics/dashboard/index.

[23] CHRIS B, RICHARD C. The D2RQ platform-accessing relational databases as virtual RDF graphs[EB/OL]. http://d2rq.org/.

[24] Free University of Bozen-Bolzano. Ontop[EB/OL]. https://ontop-vkg.org/.

[25] Oracle. Oracle big data spatial and graph[EB/OL]. https://www.oracle.com/database/technologies/bigdata-spatialandgraph.html.

(责任编辑:尹晨茹)