

“中国古代可移动文物概念参考模型”构建实践

叶祎珮

摘 要 研究旨在实现文化遗产机构之间的信息交换与集成,并提供机器可读的信息资源以满足生成式人工智能在基于事实的判断推理与辅助决策等智能应用方面的需求。故宫博物院围绕中国古代可移动文物展开研究,以 CIDOC-CRM 框架为基础,结合自上而下和自下而上的方法构建了“中国古代可移动文物概念参考模型”(CRM-ACA)。该模型也是第一款中国博物馆落地应用的本体模型。在该本体模型的基础上构建知识图谱,为故宫博物院“数字文物库”线上藏品平台提供数据服务,验证了 CRM-ACA 具有可操作性。这一研发填补了博物馆行业的空白,有助于实现博物馆藏品信息的交换集成与智能化处理。

关键词 故宫博物院; 本体模型; 概念参考模型; 信息组织; 知识图谱; CIDOC-CRM; CRM-ACA

分类号 G203

作者简介 叶祎珮,故宫博物院副研究馆员,Email:yeyipei@dpm.org.cn。

0 前言

博物馆藏品编目是管理博物馆核心资源——藏品的重要方式,也是学术研究、展览展示、宣传教育、出版等其他业务活动的基础。藏品编目工作由博物馆的专业研究人员执行,旨在对藏品的物理外观进行客观而详细的记录,并对其蕴含的艺术史、自然历史、人类学、科学等相关信息做出诠释,因此编目信息也是博物馆重要的信息资源。自 20 世纪 90 年代中期博物馆数字化浪潮掀起,藏品编目的载体由原先的纸质逐步过渡成信息系统,研究人员能够在编目系统中录入和维护编目信息,使编目信息成为互联网开放共享的资源。

随着大数据、人工智能技术不断发展,应用领域对于数据和信息的标准化也越来越重视,尤其是机器可读的信息资源,以满足生成式人工智能做基于事实的判断推理与辅助决策等智能应用的需求。为此,2019 年国

基金项目:国家重点研发计划“大型综合性博物馆数字孪生关键技术研究与示范”(2022YFF0904300)阶段性成果之一。

际标准化组织(ISO)提出了“机器可用、可读、可迁移标准”(Standards Machine Applicable, Readable and Transferable, SMART),将标准的演进划分为:纸质、开放数字格式、机器可读文档、机器可读内容、机器可释内容。

为了实现文化遗产机构之间的信息交换与融合,国际文献工作委员会(International Committee for Documentation, CIDOC)设计了一款面向文化遗产领域的本体(Ontology),即“概念参考模型”(Conceptual Reference Model, CIDOC-CRM),用于促进博物馆、档案馆和图书馆中大量分散、不兼容的文化遗产信息源基于共同的概念基础进行比较映射、数据转换等,以实现信息交换和集成。同时,通过本体表达的知识能实现“人—机”“机—机”的理解和使用,实现机器可读和互操作。然而,对于作为顶层本体的 CIDOC-CRM,在应用过程中需要根据应用场景和信息源特征对其进行适当地拓展,使其便于实际应用。

鉴于国内现有博物馆馆藏编目信息组织方式无法满足机器可读的应用需求,故宫博物院针对中国古代可移动文物编目信息特征拓展了 CIDOC-CRM,构建“中国古代可移动文物概念参考模型”(Ancient Chinese Artifacts Conceptual Reference Model, CRM-ACA),用于表示与中国古代可移动文物相关的知识内容。本文将对模型的设计思路、构建过程、可行性验证过程和实施效果一一进行阐释。

1 文献综述

从 1964 年起,国家文物局就对馆藏文物的登记账及编目卡内容做了明确规定,制定了统一的账本格式和编目卡格式,远早于“都柏林核心元数据集”或其他相关元数据标准的发布。这些元数据是为手工管理设立的,从信息规范和合理性的角度看,设计并不严谨。^[1]但是它们已经指导了全国几十年的文物工作,完全将其推翻并让全国文物管理工作从头开始是不现实的。因此,故宫博物院在建设藏品编目系统之初,是在保留部分元数据的基础上加以调整,以适应信息管理需求的。并且针对故宫藏品的信息资源特征,考虑到藏品研究和管理人员的实际业务需求设计了编目元数据。这种情况普遍存在于国内博物馆信息化建设过程中,导致博物馆之间藏品元数据存在差异,不同机构的藏品信息很难交换或集成。

按照 2018 年国家文物局《关于加强可移动文物预防性保护和数字化保护利用工作的通知》,可移动文物数字化保护利用工作即“文物信息资源数字化”包括知识图谱建设、文物信息动态著录与数据交换,“数字化服务”包括人工智能服务、博物馆智能查询、文物大数据共享服务等。对于解决异构信息源之间信息交换的问题,将元数据映射到标准本体至关重要,因为本体可作为一种中介,促进对概念及其语义关系的共同理解。CIDOC-CRM 作为一款促进文化遗产之间信息交换的本体模型,对文化遗产领域涉及的显性概念、隐性概念及两者关系提供了定义和形式化的描述,在不同信息源之间建立起共同的语言。2006 年,它被认定为国际标准 ISO 21127。2019 年,我国引入此项国际标准,发布为国家标准 GB/T 37965-2019《信息与文献 文化遗产信息交换的参考本体》。

此外,随着大数据和人工智能技术的飞速发展,越来越多的文化遗产机构开始考虑馆藏信息的更多应用前景,而计算机读取和处理数据与信息的前提是机器可读,若数据与信息采用结构化的方式来组织和存储,计算机就能够自动获取和处理它们,进而实现高效的信息管理和智能化应用。采用本体即可有效地解决这一问题,CIDOC-CRM 是基于对象的语义模型,方便转化成 RDF 或 OWL 等机器可读格式。

综上所述,采用 CIDOC-CRM 对馆藏文物编目信息进行重构不仅能够实现不同文化遗产机构藏品信息源之间的信息交换与集成,还可以为未来智能化应用打下良好的数据基础。

CIDOC-CRM 是现实世界一切实体的框架,被称为“CRMbase”的顶层本体。换句话说,CIDOC-CRM 作为一个通用框架,提供了一个用于描述不同类型实体的基本数据模型和关系。因此,不同的学科或领域可以基于 CIDOC-CRM 进行定制化的拓展,以适应特定的应用需求,而无需从头设计一个全新的本体。本体复用也是本体工程中被广泛使用和推荐的方法之一。CIDOC 提供了一系列适用于不同需求的本体模型,每个都作为 CIDOC-CRM 的拓展为各种不同类型的文化遗产管理需求提供参考。例如,FRBRoo 和 PRESSoo 用于表示文献

目录信息的语义,后者侧重于连续出版物的特殊语义;CRMInf 用于集成描述论证和推理的元数据;CRMArchaeo 用于记录考古发掘过程和相关实体及活动;CRMsci 用于描述科学领域的信息对象和活动;CRMgeo 用于描述时序实体、持续性实体以及地理信息的时空属性;CRMdig 用于记录数字化产品的生产步骤和方法;CRMba 可以描述建筑物和物质之间的关系和演化;CRMtex 可以支持古文献手抄本研究 and 记录;CRMsoc 用于描述社会现象、社会参与成员属性;CRMact 用于描述活动计划,各种活动和行为之间的关系和演变。但目前尚无一款适用于中国古代可移动文物信息组织的拓展本体,有鉴于此,故宫博物院基于 CIDOC-CRM 研发了“中国古代可移动文物概念参考模型”(以下简称 CRM-ACA),并编写配套受控词表,构建文物知识图谱,验证了 CRM-ACA 具有可操作性和可行性。

2 CRM-ACA 设计思路与研发过程

2.1 CRM-ACA 的领域

CRM-ACA 是一款概念本体,旨在服务于任何一家拥有中国古代可移动文物并将之作为藏品的文化遗产机构,它基于 CIDOC-CRM 拓展,适用于编目信息组织。就 CRM-ACA 命名而言,它包含关键词“古代中国”“文物”及“可移动文物”。以下对关键词加以说明,以限定 CRM-ACA 的适用范围。

(1)“古代中国”

根据中国史学会主持、翦伯赞、邵循正、胡华编写的《中国历史概要》一书,“古代的中国”包括原始社会、奴隶社会、封建社会,时间范围从远古一直到鸦片战争前。而将鸦片战争作为“中国古代史”下限的观点,在“中华人民共和国中央人民政府”门户网站的“国情”之古代史目录“统一的多民族国家和封建制度的逐渐衰落——明、清(鸦片战争以前)”也有对应^[2]。

(2)“文物”

根据《中华人民共和国文物保护法》总则第二条规定,“文物”的内涵包括:

- ①具有历史、艺术、科学价值的古文化遗址、古墓葬、古建筑、石窟寺和石刻、壁画;
- ②与重大历史事件、革命运动或者著名人物有关的以及具有重要纪念意义、教育意义或者史料价值的近代现代重要史迹、实物、代表性建筑;
- ③历史上各时代珍贵的艺术品、工艺美术品;
- ④历史上各时代重要的文献资料以及具有历史、艺术、科学价值的手稿和图书资料等;
- ⑤反映历史上各时代、各民族社会制度、社会生产、社会生活的代表性实物。

其中①列入“不可移动文物”范畴,但体量较小的石刻、剥离的壁画碎片也可以成为可移动文物,②不属于“古代中国”时间范畴。因此,CRM-ACA 适用于③④⑤条款列举的对象,以及体量较小的石刻和剥离的壁画碎片。

(3)“可移动文物”

“可移动文物”或“可移动文化遗产”概念在各个国家界定并不相同。国际古迹遗址理事会(ICOMOS)将“可移动文化遗产”定义为“所有的可移动的物品。它们是人类制造的、自然进化的表达和证据,具有考古、历史、艺术、科学、技术的价值和作用”。由此可见,“可移动文物”不局限于人类制造的实物,如化石也囊括在内。《中华人民共和国文物保护法》总则将“可移动文物”描述为“历史上各时代重要实物、艺术品、文献、手稿、图书资料、代表性实物等”,并且总则第二条规定“具有科学价值的古脊椎动物化石和古人类化石同文物一样受国家保护”,由此可判断在我国“化石”与“文物”有概念上的区别。<https://www.rdfyb.com/>

综上所述,“可移动文物”概念的争议可能在于是否为“人类制造”,尤其是“化石”“标本”类藏品是否属于“文物”。“化石”和“标本”是博物馆重要的藏品类型之一,从形成时间来看,“化石”和“标本”的形成时间可

能超过“古代中国”的时间范围,但 CRM-ACA 认为有必要将其纳入信息组织的对象,将“古生物化石”和“生物标本”设置为“E20 生物对象”的两个子类。考虑到“生物标本”在制作过程中可能有人类行为参与,“生物标本”也同时被设置为“E22 人造对象”的子类,即其继承 E20 和 E22 两个超类的属性。CRM-ACA 意在将“古生物化石”和“生物标本”两个类作为接口,便于自然博物馆能在 CRM-ACA 的成果基础上继续拓展。

2.2 CRM-ACA 的设计思路

CRM-ACA 构建步骤采用“斯坦福七步法”^[3],其步骤一“确定本体的领域与范围”、步骤二“考虑复用现有本体”已在前文中阐述完毕。本节从步骤三“列举本体中的重要术语”开始阐述。

2.2.1 列举重要术语

本步骤的任务是尽可能全面地收集关于“中国古代可移动文物”主题的相关术语,使主题通过具体的术语得到说明和解释,帮助推动后续步骤开展。相关术语既可能是“龙纹”“粉彩”“釉上彩”“缂丝”“青铜器”“漩涡纹”等藏品编目描述性术语,也可能是“文物分类”“创作者”“时代”“题跋”“纹饰”“质地”“尺寸”“工艺”“收藏地点”等元数据层面的术语。

此步骤暂不考虑术语之间的语义关系,如“龙纹”和“漩涡纹”都属于“纹饰”,“粉彩”和“釉上彩”都属于“工艺”,“穿花龙纹”与“龙穿花纹”是同义词等,只是尽可能多地收集术语,保证术语的全面性。CRM-ACA 的术语来源如下。

(1)《故宫博物院文物详细编目则例》^①。该则例是故宫博物院器物部、宫廷历史部、书画部、图书馆等多个部门的研究人员合力编写的编目规范,能够较全面地反映中国古代可移动文物情况。它包括元数据和描述性术语。

(2)盖蒂研究所《艺术品描述类目》(*Categories for the Description of Works of Art*, CDWA)。CDWA 由美国 Getty 基金会(J. Paul Getty Trust)和学院艺术协会(College Art Association)联合资助的艺术信息工作组 AITF (the Art Information Task Force)制定,是为从事艺术史研究、艺术品管理的人员以及信息技术专家设计的元数据标准,用来描述艺术品,主要在博物馆使用。该标准在国际上具有广泛影响力,侧重元数据术语。

(3)故宫博物院“数字文物库”用户检索记录。“数字文物库”是故宫博物院重要的藏品检索平台,自 2019 年 6 月发布至 2023 年 5 月,累计独立访客数量达到 4006456 次,累计检索超过 566 万次,能直观反映用户检索行为。用户检索记录侧重描述性术语,它与《故宫博物院文物详细编目则例》结合使用能使术语既覆盖领域专家的专业术语,如“宫衣”“武生巾”,又覆盖用户使用的日常表达,如“衣服”“帽子”。

(4)行业参考文献,包括但不限于:中央编译出版社《中国文物大辞典》、上海辞书出版社《中国考古学大辞典》、故宫出版社《故宫辞典》、中华书局《中国历代职官别名大辞典》等。

CRM-ACA 在步骤三共计收集术语五万余个。

2.2.2 定义类及其层次结构

“类”(Class)是一组具有一个或多个共同特征的事物或概念聚合而成的类别,该特征能作为判别标准把这一类事物或概念识别出来。CIDOC-CRM 鼓励“通过类型的层级结构”对模型进行拓展。CRM-ACA 采用自上而下与自下而上相结合的方式构建类的层级(图 1):元数据级别的术语适合采用自上而下的方式形成本体草稿 A,描述性术语适合采用自下而上的方式形成本体草稿 B,将 A 与 B 进行对比、改善,使二者衔接融合形成完整的类的层级,再与 CIDOC-CRM 末端的类对齐。

自上而下的构建过程以《故宫博物院文物详细编目则例》和 CDWA 的部分元数据术语作为演示,主要工作是判断元数据术语在本体中的应用,如“类”“关系”或“属性”(表 1)。

根据语义场理论,若干个词语可能具有共同义素的义位,存在语义层次、语义范围重叠等现象。以上术语

① 内部资料,尚未出版。

彼此之间可能存在此类现象,需进一步分析其的语义聚合关系,为本体梳理出概念界限相对清晰的“类”并合理归入 CIDOC-CRM 的框架内,与既有“类”形成“子类属于一个超类”(isSubclassOf)关系。

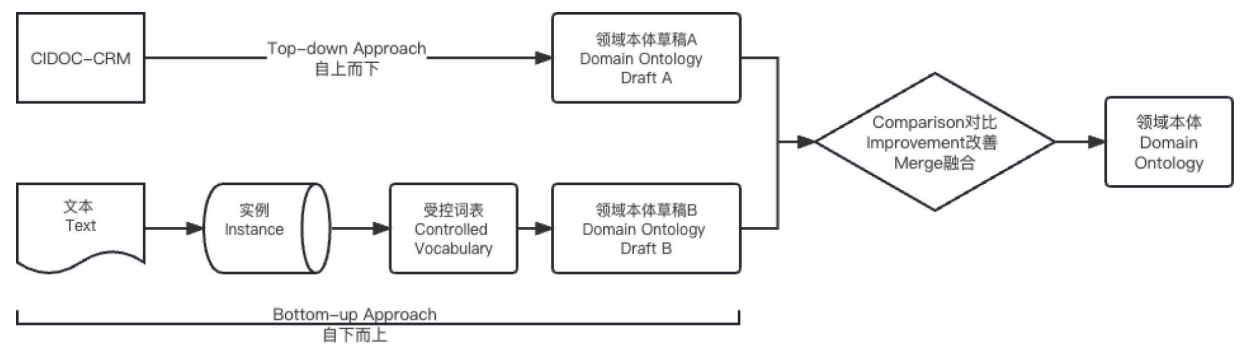


图 1 CRM-ACA 类的层级构建方法

本节将“时代”“时期”“年号”“地缘政治单位(政权)”“朝代”“文化类型”“风格”作为一组术语进行语义辨析。

- (1)时代:社会发展的不同阶段或时期,如原始时代、社会主义时代。
- (2)时期:发展过程中的一段时间,如全盛时期、冰河时期、孩童时期。
- (3)年号:皇帝纪年的名号。一个皇帝在位期间可用几个年号,但明、清两代,一个皇帝一般只用一个年号。
- (4)地缘政治单位(政权):拥有共同的语言、文化、种族、宗教、领土、政权或者历史的社会群体;从狭义的角度,它是一定范围内的人群所形成的共同体形式。
- (5)朝代:建立国号的君主(一代或若干代相传)统治的整个时期。
- (6)文化类型:指考古学文化,几种特定型式的器物,经常在一定地区的某一类型的住址或墓葬中共同出土,这种具有特定组合关系的遗存,便属于一种考古学文化。
- (7)风格:一个时代、一个民族、一个流派或一个人的文艺作品所表现的主要的思想特点和艺术特点。

根据 CIDOC-CRM 之“E4 时期”的范围说明 (Scope Note)^①,以上七个术语均属于 E4 的概念范畴。其中 (1)(2)与 E4 的概念基本重叠。(5)的概念根据《现代汉语词典》可分两种:某一世系帝王的统治时期;某一帝王的统治时期。详细辨析见表 2。

词义 1 可与“地缘政治单位(政权)”概念合并,采用术语“地缘政治单位”以免与词义 2 概念混淆。词义 2 之情况则需分别处理,作为“E4 时期”的实例(Instance)或创建“年号”类作为“地缘政治单位”的子类。

另外,年号纪年是中国古代纪年的重要方法之一,此处对“年号纪年”与“年号”“纪年”做语义辨析,详见表 3。

综上所述,CRM-ACA 以自上而下的构建方式将 CIDOC-CRM 之“E4 时期”拓展如表 4。

① “E4 时期”包括在时间和空间中发生的一系列连贯的现象或文化表现。它的实例是这些现象的社会性或物质性的连贯特征,而不是相关时空范围。不同的时期可以在时间和空间上重叠和共存,比如一个游牧文化与一个定居文化同时存在于同一地区和时间。通常,本类用于描述史前时期或“新石器时代”“明代”或“麦卡锡时代”等历史时期,但也将地缘政治单位和聚落活动视为“E4 时期”的特殊情况,但不相关现象的规模大小做任何假设,特别是所有事件均被视为由连贯现象组成的综合过程,因此“E4 时期”是“E5 事件”的超类。文化遗产编目使用该来记录文化和艺术时期。对于“艺术风格”,有两种不同的概念化方式,一种是基于物理特征,另一种是基于历史背景。例如,“印象派”可以被视为欧洲影响力范围内的一个时期,大约从 1870 年到 1905 年,在这个时期内由包括莫奈、雷诺阿、皮萨罗、西斯利和德加等艺术家创作了具有特定特征的绘画作品。或者它也可以被看作适用于所有具有印象派画家作品特点的绘画,而不考虑历史背景。第一种解释是“E4 时期”的实例,第二种解释是“E5 类型”的实例,定义的是形态学对象类型。另一个“E4 时期”的特定情况是由其物理痕迹所证明的一系列活动和现象的实际范围,这些痕迹定义了一个定居点,例如尼尼微城的人口聚集时期。

表 1 元数据术语梳理示范

术语	定义	类型	应用于 CRM-ACA		
			类	关系	属性
文物名称 (来源《故宫博物院文物详细编目则例》)	博物馆藏品名称	自由文本			E22 人造对象
细类 (来源《故宫博物院文物详细编目则例》)	博物馆根据文物属性做出的文物分类与细分类	受控词表	E55 类型→ 藏品类型		
级别 (来源《故宫博物院文物详细编目则例》)	文物定级参照国家文物定级序列,即一级、二级、三级、一般文物	受控词表	E55 类型→ 级别		
时代 (来源《故宫博物院文物详细编目则例》)	指代文物创造年代,有若干种描述形式。例如,公元纪年(1906 年),世纪年代(20 世纪 40 年代),具有延续性的政权(唐/明/清),年号(雍正),年号纪年(贞观二年),历史时段(清早期/19 世纪中叶)等	受控词表	E4 时期		
作者 (来源《故宫博物院文物详细编目则例》)	文物的创作者	受控词表		组合式语义关系 E22 人造对象 — P108i 被产生经由→E12 产生 E12 产生 — P14 的实施者是→E21 人物	
styles/periods/ groups/movements (来源 CDWA)	A description of a work of art that associates it with a defined style, historical period, group, school, or movement whose characteristics are represented in the work (对艺术作品的描述,将其与作品中表现出的特定风格、历史时期、群体、学校或运动相关联)	受控词表	E4 时期		
object/work type (来源 CDWA)	The kind of object or work described (所描述的对象或作品的种类)	受控词表	E55 类型→ 物件类型		

由此自上而下构建方法形成本体草稿 A。而自下而上构建过程首先需要将零散的术语根据语义做聚合,梳理术语的语义关系。为了灵活而有效地组织术语并进行管理,且避免本体构建工具 Protégé 使用中实例名称不可重复而导致无法处理同形多义词带来的不便,CRM-ACA 采用叙词表作为自下而上梳理术语的工具。

表 2 “朝代”词义辨析

“朝代”词义辨析	词义	情况	示例	处理方案
	词义 1： 某一世系帝王的统治时期			秦/唐/宋/明 并入“地缘政治单位”类
	词义 2： 某一帝王的统治时期	当帝王无年号	秦始皇统治时期/ 汉武帝统治时期	作为“E4 时期”的实例
		当一位帝王有多个年号	武则天统治时期	作为“E4 时期”的实例
			天授/长寿/神功	创建“年号”类
		当一位帝王仅有一个年号,多 见于明、清两代,今人用某个 年号指代某帝王的统治时期	康熙帝统治时期/ 雍正帝统治时期	作为“E4 时期”的实例
			康熙/雍正	创建“年号”类

表 3 “年号纪年”“年号”“纪年”语义辨析

术语	情况	示例	处理方案
年号纪年	“年号纪年”是“年号”产生后使用的一种纪年方式	开宝二年/洪武元年/乾隆三十年	创建“年号纪年”类,作为“年号”的子类。“年号”与“年号纪年”形成“P9 包括时期”的语义关系
年号	帝王用来纪年的一种名号	开宝/洪武/乾隆	创建“年号”类,作为“地缘政治单位”的子类。“年号”与“年号纪年”形成“P9 包括时期”的语义关系
纪年	在“年号”产生前,古代中国使用的主要纪年方式有:王位序数纪年、大事纪年、太岁纪年、干支纪年等	惟王十祀又五彤日/ 大司馬昭陽敗晉師於襄陵之歲/歲在涪灘/甲午	1. 创建“古代纪年”以区分于“公元纪年” 2. 创建“王位序数纪年”“太岁纪年”“干支纪年”,作为“古代纪年”的子类

表 4 CRM-ACA 对 CIDOC-CRM 之“E4 时期”的扩展

超类	isSuperclassOf	isSuperclassOf	isSuperclassOf
E4 时期 Period	地缘政治单位 Geopolitical Unit	年号 Reign Title	年号纪年 Year by Reign Title
	文化类型 Archaeological Culture		
	风格 Style		
	古代纪年 Year by Ancient Time-reckoning	王位序数纪年 Regnal Year	
		太岁纪年 Year by Taisui	
		干支纪年 Year by Sexagenary Cycle	
		年号纪年 Year by Reign Title	

根据国家社科基金项目“基于本体和知识集成实现中文叙词表的升级、共享和动态完善”的研究成果,叙词表与本体二者在定义表述上都是“通过受控词汇来表达概念的概念系统,都提供了对领域知识的共同理解与描述,都追求概念及其之间关系的明确化(explicit)和描述的形式化(formal)”。^[4]且词表本质上是一种具有等级关系的知识组织形式,与本体模型中关系类型之一的“超类—子类”概念关系(sub-super-concept)^[5]同属于树形结构。因此,从理论上,词表的“IsA”语义关系可以转化成本体的“子类属于一个超类”的“isSubclassOf”关系。

自下而上构建过程使用受控词表将零散的术语组织成层级结构,以“数字文物库”用户检索记录和行业参考文献的部分术语作为演示(表5)。

表 5	术语聚合演示
原始术语	术语聚合
康熙、红色、青铜尊、高足杯、乾隆、花盆、植物图案、扇子、如意、雍正、蓝色、梅瓶、博山炉、玉壶春瓶、龙纹、嘉庆、黄色、首饰、衣服、仙鹤……	(年号)康熙、乾隆、雍正、嘉庆……
	(颜色)红色、蓝色、黄色……
	(物件类型)青铜尊、高足杯、花盆、梅瓶、博山炉、玉壶春瓶、扇子、如意、首饰、衣服……
	(图案与纹样)植物图案、龙纹、仙鹤……

术语聚合后,再使用合适的高维度术语统称某一类的术语,如“青铜尊”“高足杯”“梅瓶”等可以使用“物件类型”统称它们。零散的术语需要分组做语义辨析,如“青铜尊”是依材料区分的“尊”,“高足杯”是依形式区分的“杯”,“花盆”是依功能区分的“盆”。按照逻辑,“青铜尊”的同层级术语还应包括“珐琅尊”“玻璃尊”等,“依形式区分的尊”包括“牺尊”“象尊”“天鸡尊”“石榴尊”等,而“依材料区分的尊”和“依形式区分的尊”都属于“尊”的下位词,由此构成层级关系。位于层级最顶部的术语成为“类”的候选词,比如“物件类型”作为该类术语聚合后最具概括性的术语将被创建成一个新类“物件类型”,并且作为 CIDOC-CRM“E55 类型”的子类,这是本体草稿 A 和 B 上下对齐融合的过程。CRM-ACA 基于 CIDOC-CRM 拓展了 36 个类,每个类有一个对应的受控词表。

2.2.3 定义类的属性

CIDOC-CRM 文档将两个“类”之间的关系称为“属性”(Property),为避免概念混淆,本文将其称为“关系”,而将描述“类”中实例(Instance)的特征或特性称为“属性”(Slot)。定义属性大致有以下两种方式。(1)将“类”的内在特征和外在特征设置成“类”的属性。例如将某件藏品的编目信息作为它的属性值,如藏品 a 有属性“纹饰”、值为“漩涡纹”,有属性“材质”、值为“陶”。(2)把具有特殊性或唯一性的标识设为属性,例如,藏品 a 有属性“藏品名称”、值为“斗彩龙凤穿花纹带盖梅瓶”,有属性“藏品编号”、值为“故 00152086”。CRM-ACA 认为具有共性的信息,即(1)更适合作为实例,与它相关的其他实例通过“关系”关联。如将“漩涡纹”作为“图案与纹样”类的一个实例,藏品 a 经关系“P62 描绘”关联“漩涡纹”。

2.2.4 创建实例

CRM-ACA 的自下而上构建过程产出了 36 个受控词表,词表最顶部的术语纳入本体“类”,其他术语成为该“类”的实例,实例之间的关系依据 CIDOC-CRM 定义的关系进行命名。在 CRM-ACA 中,“物件类型”是“E55 类型”的子类,“物件类型”词表的术语成为该类的实例,术语的层级关系采用“P127 有上位词”表达。而“E4 时期”的子类“地缘政治单位”与“年号”的实例则应采用“P9 包括时期”表达,如“清”“包括时期”“雍正”。目前 CRM-ACA 建设实例近三万个,经过故宫博物院器物部、书画部、宫廷历史部、考古部、图书馆在内的十几位领域专家审校,较具有学术准确性。

3 CRM-ACA 可行性验证

为面向智能检索、语义问答、辅助大数据分析和推理判断等可能的人工智能应用场景,故宫博物院选择知

识图谱实践基于 CRM-ACA 的数据建设。知识图谱能够通过节点之间的相互关联来表达真实世界的知识,节点之间的关系是有向的、多层次的,能够表达现实世界中非常复杂的语义,有助于更全面地理解和挖掘知识并从中获得洞察和发现新的知识。

本节通过演示将传统藏品编目形式转换为基于 CRM-ACA 的信息组织形式的方法——以实例为节点、以关系为边重构馆藏编目信息(表 6),验证了 CRM-ACA 是可以满足中国古代可移动文物编目需要的。

表 6 编目信息在 CRM-ACA 中的表达

编目元数据	值	基于 CRM-ACA 表示
藏品编号	故 00162541-6/67	故 00162541-6/67_被标识为(is identified by)_故 00162541-6/67
藏品名称	黄地粉彩梅鹊图碗	故 00162541-6/67_有标题(has title)_黄地粉彩梅鹊图碗
数量	1	故 00162541-6/67_的组件数量是(has number of parts)_1
计量单位	件	
类别	陶瓷	故 00162541-6/67_有类型(has type)_陶瓷
细类	瓷器	故 00162541-6/67_有类型(has type)_瓷器
题材	喜上眉梢	故 00162541-6/67_描绘(depicts)_喜上眉梢
级别	一级	故 00162541-6/67_有类型(has type)_一级品
时代	清 同治	故 00162541-6/67_被产生经由(was produced by)_故 00162541-6/67 的产生 故 00162541-6/67 的产生_属于时空集合(falls within)_同治 故 00162541-6/67 的产生_发生时段是(has time-span)_1862-1874
质地	瓷	故 00162541-6/67_的材质是(consists of)_瓷
纹饰	梅鹊	故 00162541-6/67_描绘(depicts)_梅鹊
尺寸	高 9.3 厘米	故 00162541-6/67_的度量规格是(has dimension)_故 00162541-6/67 的高度 故 00162541-6/67 的高度_有数值(has value)_9.3
款识	同治年制	故 00162541-6/67_承载(carries)_同治年制
工艺特点	粉彩,黄釉	故 00162541-6/67 的产生_使用通用技术(used general technique)_粉彩 故 00162541-6/67 的产生_使用通用技术(used general technique)_黄釉

知识图谱的关联性还具有良好的可扩展性,通过不断地添加新实体、新关系能反映更新的知识。通过建立与特定实例“故 00162541-6/67”的直接关联,可以实现“故 00162541-6/67”与更多实例之间的间接关系。例如,在“图案与纹样”类中,“梅鹊”同时为“梅”和“喜鹊”的下位词,“梅”和“喜鹊”又分别为“花卉”和“鸟”的下位词,层层关联使实例“故 00162541-6/67”能够与“花卉”和“鸟”等更广泛的节点产生间接关系,丰富了实例“故 00162541-6/67”的语义连接,为进一步的语义推理和知识发现提供更多可能。知识图谱这一数据关联优势已在故宫博物院的“数字文物库”平台中有所体现(图 2)。2023 年 2 月,基于 CRM-ACA 建设的知识图谱上线,开始为故宫博物院“数字文物库”提供藏品检索数据服务。基于知识图谱的藏品检索突破了传统的关键词匹配检索技术瓶颈,更加智能,丰富了用户的检索体验。例如“动物图案”“鸟(动物)”“植物纹样”等词能够检索到“故 00162541-6/67”,“唐卡”“瓷瓶”“衣服”等过去无法检索到藏品的语句现在能够有效地找到相应的藏品。

此外,基于 CRM-ACA 的知识图谱还有助于研究者发现藏品之间的深层关系。比如“故 00162541-6/67”是为同治皇帝大婚烧制的瓷器套组之一,烧制样式可见于“内府色绘黄地粉彩梅鹊图碗纸样”,该纸样还绘有同是“黄地粉彩梅鹊图”的盒、渣斗、盘、碟、匙等其他器物形制,实物同样收藏于故宫博物院(图 3)。

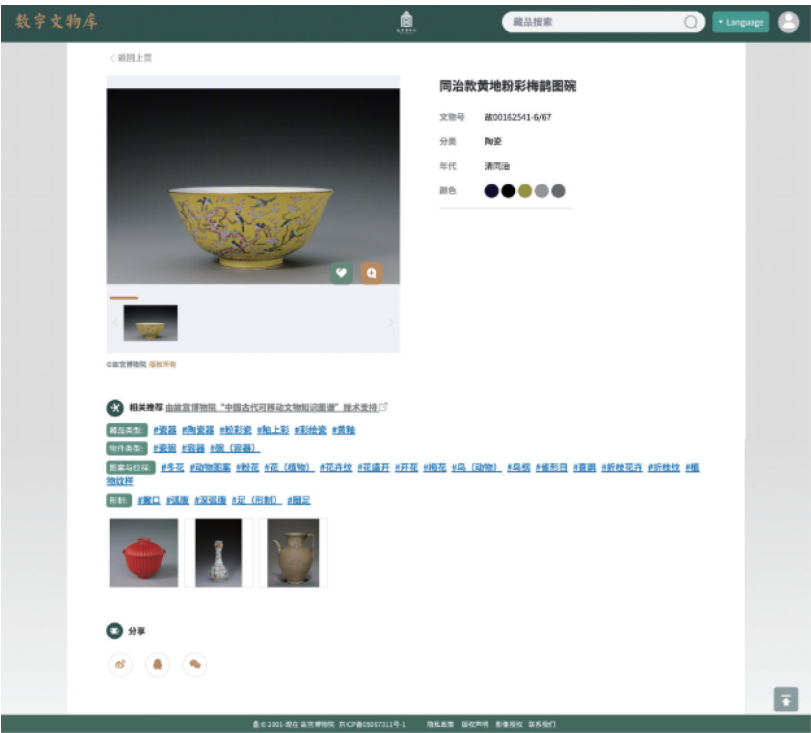


图2 知识图谱应用于“数字文物库”智能检索

以上事实可表达如下。

故 00162541-6/67_记载于(is documented in)_内府色绘黄地粉彩梅鹊图纸样

故 00156178_记载于(is documented in)_内府色绘黄地粉彩梅鹊图纸样

故 00162541-6/67_相似于(shows features of)_故 00156178

“故 00162541-6/67”示例演示了 CRM-ACA 在中国古代可移动文物信息组织领域的能力和潜力。经统计,仅围绕“藏品”(“E22 人造对象”),CRM-ACA 就能表示其与其他 42 个“类”存在的 63 种语义关系(表 7)。尽管“E22 人造对象”与一些类,如“E52 时段”,没有直接关系,但仍然可以通过组合式语义来表示相关信息。这意味着即使没有直接关联,也可以通过多关系的组合来间接表示相关信息。



图3 与“故 00162541-6/67”相关的其他藏品

表 7 “E22 人造对象”作为定义域和值域的关系列表

E22 作为定义域			E22 作为值域		
E22	P1 is identified by	E41 Appellation	E22	P8i witnessed	E4 Period
E22	P2 has type	E55 Type	E22	P12i was present at	E5 Event
E22	P3 has note	E62 String	E22	P13i was destroyed by	E6 Destruction
E22	P43 has dimension	E54 Dimension	E22	P15i influenced	E7 Activity
E22	P44 has condition	E3 Condition State	E22	P16i was used for	E7 Activity
E22	P45 consists of	E57 Material	E22	P17i motivated	E7 Activity
E22	P46 is composed of	E18 Physical Thing	E22	P19i was made for	E7 Activity
E22	P48 has preferred identifier	E42 Identifier	E22	P24i changed ownership through	E8 Acquisition
E22	P49 has former or current keeper	E39 Actor	E22	P25i moved by	E9 Move
E22	P50 has current keeper	E39 Actor	E22	P30i custody transferred through	E10 Transfer of Custody
E22	P51 has former or current owner	E39 Actor	E22	P31i was modified by	E11 Modification
E22	P52 has current owner	E39 Actor	E22	P34i was assessed by	E14 Condition Assessment
E22	P53 has former or current location	E53 Place	E22	P39i was measured by	E16 Measurement
E22	P54 has current permanent location	E53 Place	E22	P41i was classified by	E17 Type Assignment
E22	P55 has current location	E53 Place	E22	P67i is referred to by	E89 Propositional Object
E22	P56 bears feature	E26 Physical Feature	E22	P70i is documented in	E31 Document
E22	P57 has number of parts	E60 Number	E22	P71i is listed in	E32 Authority Document
E22	P59 has section	E53 Place	E22	P92i was brought into existence by	E63 Beginning of Existence
E22	P62 depicts	E1 CRM Entity	E22	P93i was taken out of existence by	E64 End of Existence
E22	P65 shows visual item	E36 Visual Item	E22	P108i was produced by	E12 Production
E22	P101 had as general use	E55 Type	E22	P110i was augmented by	E79 Part Addition
E22	P102 has title	E35 Title	E22	P111i was added by	E79 Part Addition
E22	P103 was intended for	E55 Type	E22	P112i was diminished by	E80 Part Removal
E22	P104 is subject to	E30 Right	E22	P113i was removed by	E80 Part Removal
E22	P105 right held by	E39 Actor	E22	P123i resulted from	E81 Transformation
E22	P128 carries	E90 Symbolic Object	E22	P124i was transformed by	E81 Transformation
E22	P130 shows features of	E70 Thing	E22	P129i is subject of	E89 Propositional Object
E22	P137 exemplifies	E55 Type	E22	P136i supported type creation	E83 Type Creation
E22	P156 occupies	E53 Place	E22	P138i has representation	E36 Visual Item
E22	P196 defines	E92 Spacetime Volume	E22	P157i provides reference space for	E53 Place
			E22	P188i is production tool for	E99 Product Type
			E22	P195i had presence	E93 Presence
			E22	P198i is held or supported by	E18 Physical Thing

4 结语

CRM-ACA 是基于 CIDOC-CRM 框架拓展的本体模型,旨在服务于中国古代可移动文物的信息组织。本文详细介绍了 CRM-ACA 的设计思路、研发步骤和构建过程,并验证了其在实践中的可行性。通过应用 CRM-ACA,博物馆能够有效地组织或重组馆藏文物的信息,使其变得机器可读,更好地利用和分析文物数据,从中挖掘出更多的知识和价值。这不仅有助于提高文物管理和研究的效率,也为博物馆面向大数据和人工智能时代的新应用挑战提供了重要的支持。

值得强调的是,作为一个动态发展的本体模型,CRM-ACA 在未来将继续丰富更多语义关系,并且有计划面向不可移动文物继续拓展,为文物研究和保护提供更为丰富和多样的视角和方法。

综上所述,CRM-ACA 作为第一款中国博物馆落地应用的本体模型,展现了其在博物馆信息管理和面向新应用挑战方面的潜力,同时也为未来拓展到不可移动文物领域提供了动力,期待更多文化遗产机构共同参与使用。

参考文献

[1] 胡锤,梅雪. 故宫文物管理信息系统基础数据表的建立 [EB/OL]. (2016-08-04) [2023-06]. https://www.dpm.org.cn/study_detail/100197.html.
[2] 中华人民共和国中央人民政府. 古代史 [EB/OL]. (2020-10-29) [2023-09]. https://www.gov.cn/guoqing/2020-10/29/content_5555760.htm.
[3] NOY N F, MCGUINNESS D L. Ontology development 101 :a guide to creating your first ontology [EB/OL]. (2019-09-06) [2023-06]. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
[4] 曾新红. 中文叙词表本体——叙词表与本体的融合 [J]. 现代图书情报技术, 2009(1): 34-43.
[5] STEVENS R. What is an ontology? [EB/OL]. (2001-07-19) [2023-06]. <http://www.cs.man.ac.uk/~stevensr/onto/node3.html>.

Ontology Construction: Ancient Chinese Artifacts Conceptual Reference Model

Ye Yipei

Abstract This study aims to achieve information exchange and integration among cultural heritage institutions and provide machine-readable information resources to meet the needs of generative artificial intelligence in factual reasoning and decision support. Focused on ancient Chinese movable artifacts, the research utilizes the CIDOC-CRM framework as a foundation and combines top-down and bottom-up approaches to construct the Ancient Chinese Artifacts Conceptual Reference Model (CRM-ACA). Building upon this ontology model, a knowledge graph is created to facilitate data services for the Palace Museum's Online Collection Platform, thereby validating the operability of CRM-ACA. This development fills gaps in the museum industry and promotes the exchange, integration, and intelligent processing of museum collection information.

Key words the Palace Museum; ontology model; conceptual reference model; information organization; knowledge graph; CIDOC-CRM; CRM-ACA