

面向人工智能的博物馆

藏品知识组织

——以故宫博物院“中国古代可移动文物概念参考模型”为例^{*}

庄 颖

内容提要 本文概述了大数据时代中国博物馆在藏品知识组织体系构建和共享方面面临的挑战，诸如文物名称复杂，信息结构化和标准化程度低，相关概念间缺乏有效关系模型等。上述问题降低了藏品信息的机器可读性，给规模化利用藏品数据集进行数字人文研究带来困难。基于此，故宫博物院构建了“中国古代可移动文物概念参考模型”（CRM-ACA），在专家编目基础上，内化领域知识体系、增强语义网络并构建知识图谱。依托故宫博物院体系完备、品类丰富的藏品，该模型为组织和利用中国古代可移动文物知识提供了一个可扩展的智能化框架，其对齐开放数据标准，为研究和利用释放了新的潜力。

关键词 概念参考模型 博物馆藏品 故宫博物院 数字人文 博物馆数字化

在一篇讨论进行知识组织工作必要性的文章中，安德烈亚·沙恩霍斯特(Andrea Scharnhorst)和理查德·P.斯密雷利亚(Richard P. Smiraglia)谈及，“正如过去三个世纪以来先辈们试图理解由印刷术释放的混沌一般，大家都踏上了理解万维网之混沌的旅程”^①。知识大爆炸的时代已经到来，如何掌握有用的知识及有效地处理、获取、整合和分析信息已成为当前亟需解决的问题。对知识的分类与排序几乎是伴随着人类知识的形成而产生的，其本身就是通过观察、归纳等方法将原始材料变成有效信息的一种智性活动。“知识组织”可谓对知识分类排序的系统化、科学化，19世纪末20世纪初欧洲的最后一次工业革命带来了知识的显著增长，图书馆在对这一时期扮演了特殊的角色，从而催生了“知识组织”作为一个专属科

* 本文为文化和旅游部2020年度文化艺术和旅游研究项目信息化发展专项“应用于人工智能搜索的可移动文物‘概念参考模型’研究”（项目编号：MCT2020XZ11）部分成果，由文化和旅游部及故宫博物院提供课题研究经费资助。

① Scharnhorst, A.; Smiraglia, R. P. (2021). “The Need for Knowledge Organization. Introduction to the book *Linking Knowledge: Linked Open Data for Knowledge Organization (Chapter 1)*”. In R. P. Smiraglia; A. Scharnhorst (Eds.), *Linking Knowledge*. Baden-Baden, Ergon –Nomos. p.13. <https://doi.org/10.5771/9783956506611-1>, 查阅时间：2023年8月1日。

学领域的兴起，这也决定了长期以来图书馆在这一领域所占据的核心地位。一般而言，传统知识组织主要面向文献、档案等，通过杜威十进分类法(DDC)、美国国会图书馆图书分类法(LCC)、国际十进分类法(UDC)在内的传统分类系统(可追溯至约1876年)，以及后来发展的包括符号学方法、“批判解释”方法等在内的其他方法^①，经由图书情报学、档案学领域的专业人士如图书管理员、档案管理员等^②，从事相关研究与实践。

信息时代的到来，尤其是过去二十年间互联网技术高速发展引发的信息爆炸式增长，让知识组织突破了原有的核心学科领域，为人们利用数字技术，增强对信息的理解和分析，乃至治理“万维网之混沌”提供了一类解决方案。与此同时，这一实践的重心亦随之发生改变。从前，知识组织往往通过人工对书籍、档案等文本的信息进行字面意义的分类、组织，为利用这些文本信息提供便利——即便这一过程后来利用计算机来完成，其行为仍是“句法”层面的^③，通过关键词逐字匹配进行信息检索便是一例。信息的爆炸式增长使得这一领域纯人工的实践不再可持续，从而驱动信息组织^④实践进入到“语义”层面，其终极目标便是要让计算机能够理解信息、数据和知识，从而借助算力的持续增长帮助人类更高效地处理信息和数据，获取知识。基于语义网络(Semantic Network)的技术无疑是当下知识组织的前沿领域，不断有新的方法、工具和手段应运而生。有了作为语义网络模型层的本体(ontology)加持，计算机可以进行不同程度的计算和推理，通过相对复杂的概念关系，实现对知识的标引和检索^⑤。

自20世纪90年代以来。随着以藏品为核心开展的业务信息系统建设及数字化的逐步推进，作为收藏、保护、研究、阐释和展示物质与非物质遗产的机构，博物馆持续积累了大量的藏品基础信息及包括数字影像、数字档案等在内的数字资源。2010年前后，如何全面开放这些信息资源，更好地服务于学术研究及公众、社区，提升资源利用效率，成为了海内外博物馆界普遍关注且充分讨论的话题。2008年欧洲数字文化遗产平台(Europeana)的发布，Open GLAM运动于2010年的发端，以及阿姆斯特丹荷兰国立博物馆(Rijksmuseum)于2016年引领的藏品数据及影像持续免费开放获取的风潮，均为博物馆馆藏文化遗产信息共享与资源开放的重要里程碑。这些节点展示了博物馆持续推进文化遗产资源开放共享的变

① Broughton, Vanda, Hansson, Joacim, Hjørland, Birger and López-Huertas, Maria J. 2005, “Knowledge organisation: Report of working group 7”. In Kajberg, L. and Lørring, L., eds., *European curriculum reflections on education in library and information science*, Royal School of Library and Information Science, Copenhagen. <http://www.db.dk/LIS-EU/workshop.asp>, 查阅时间：2023年8月1日。

② Hjørland, Birger. “What is Knowledge Organization (KO)?” in *Knowledge Organization*, 35(3/2), p.86.

③ Warner, Julian 2007. “Description and search labor for information retrieval”. *Journal of the American Society of Information Science and Technology* 58:12, pp.1783-90.

④ 本文中“知识组织”“信息组织”可换用。在知识组织领域，两个概念略有差异，一般情况下在图书情报学范畴之下“知识组织”更为常用，“信息组织”则不明确针对具体学科或领域，常见于一般性谈论信息科学的语境之中。详见Hjørland, Birger. “Is Knowledge Organization= Information Organization?” in *Categories, Contexts and Relations in Knowledge Organization*, Ergon-Verlag, 2012, pp. 8-14.

⑤ 常春《网络环境下叙词表编制与发展》，科学技术文献出版社，2015年，第131页。

革过程。

尽管在藏品信息化和数字化建设方面起步较晚，但由于社会经济的全面发展、优秀的数字化基础设施及社会需求的充分释放，中国博物馆在数字资源建设和开放整合方面近年大有作为，不断缩小与欧美同行的差距。然而，信息资源的建设、积累与全面开放、有效利用之间仍存在许多难以克服的问题，其中“明明有却找不到、不好用”的问题尤为突出。文化遗产领域数字人文研究(digital humanities)的兴起所带来的对大规模高质量数据的需求，让这一问题更加凸显。本文以故宫博物院“中国古代可移动文物概念参考模型”的构建缘起为例，从藏品知识组织的视角，梳理这一问题的由来与现状，探讨应用语义网络技术，依托文化遗产信息标准化顶层设计进行的藏品数据“富化”(enrichment)、机器可读程度提升等举措形成的一套解决方案，以期与业界同仁共谋博物馆藏品信息资源建设的深化发展。

一 数字人文研究的兴起与博物馆藏品的知识组织：现状与挑战

构成社会科学和人文科学(social sciences and humanities)的多样异构的知识领域，包括其在文化遗产中的应用，随着语义网的发展壮大，正在为链接开放数据云带来多维的丰富性^①。数字人文是计算机学科(特别是计算语言学和信息科学)与人文学科的交集，可以看作是数字技术对人文学科传统研究范式、技术、方法与理论的改造与变革^②，可谓人文研究领域的“数字化转型”。过去二十年间，关于“数字人文”的定义及其特征的论述可谓汗牛充栋、自成一体^③，其核心在于通过易于计算的建模(computationally tractable modelling)为人文学科提供一种新方法，分析大量文献和其他人造物，探知人类无法洞察的属性与特质，从而突破传统人文学科在认识论(epistemology)方面的局限^④。

数字人文研究方法具有在宽广的时空尺度上进行数据挖掘和文本分析的使用特性，人文学科数据集因而须以机器可读、互操作的结构化、规模化、标准化形式存在，它们不仅是数字人文研究的基础，更是不可或缺的研究对象。因此，近年来国内高校信息资源管理学科及博物馆数字化领域专家均非常关注作为知识“富矿”的文化遗产领域数字人文研究中的数据供给问题^⑤。馆藏文物信息资源是博物馆拥有的最基础、最核心的文化遗产数据集，这笔数据的规模和质量直接决定了博物馆对数字人文研究发展的支撑

① Scharnhorst, A.; Smiraglia, R. P., op. cit., p.13.

② Berry, David M. "Introduction: Understanding the digital humanities." In *Understanding digital humanities*, Palgrave Macmillan, London, 2012, pp.1-20.

③ Kirschenbaum, Matthew. "What is 'Digital Humanities', and why are they saying such terrible things about it?." *differences* 25, no. 1 (2014), pp.46-63.

④ Peter Boot, "Defining Digital Humanities. A Reader", in Melissa Terras, Julianne Nyhan and Edward Vanhoutte (eds.), *Digital Scholarship in the Humanities*, Volume 30, Issue 1, April 2015, pp.154-156, <https://doi.org/10.1093/lrc/fqu025>, 查阅时间：2023年8月1日。

⑤ 刘健、陈晴《数字人文与博物馆》，《中国博物馆》2018年第2期，第7页。

力度，同时也决定了在大数据时代博物馆能够在多大程度上借力推动馆藏文物广泛深入的整理、认知、研究和利用，尤其是借助语义网络等新一代知识组织工具对相关领域知识深度和广度的拓展，彰显文化遗产在公共文化服务领域所能发挥的社会效益。

我国博物馆藏品信息化、数字化的历程始于20世纪90年代，最初主要服务于机构内部管理和业务工作的需求，藏品信息组织及相关数字资源的建设、保存及管理实践普遍依托于藏品管理系统(Collection Management System，简称CMS)与数字资源管理系统(Digital Asset Management System，简称DAM)相结合的模式开展。随着博物馆数字化转型的深化，以及近年来藏品信息公开的大势所趋，博物馆官方网站成为了对外提供藏品信息资源公共文化服务的核心渠道，常见的形式是从CMS调用部分藏品编目数据、关联DAM中的藏品影像发布，以保证公布的信息与业务数据的一致性，有利于端到端的数据一致性控制，从而保障数据服务的可持续性。然而，藏品信息化、数字化后形成资源，并不代表藏品信息便直接具有了可获取性，在对藏品信息及其所蕴含的知识进行系统性组织之前，它们都是原始数据，还不是信息。从利用的角度而言，CMS等“业务驱动”的藏品信息资源的分类、描述、注释是片段式的，高度依托于从事藏品管理和研究领域专家自身已有的知识体系，缺乏与相关领域知识的规范化、体系化关联，为利用侧所进行的整合与共享造成了一定障碍，而数字人文研究兴起所带来的对机器可读的规模化文化遗产数据集的需求也让这个问题愈加凸显。

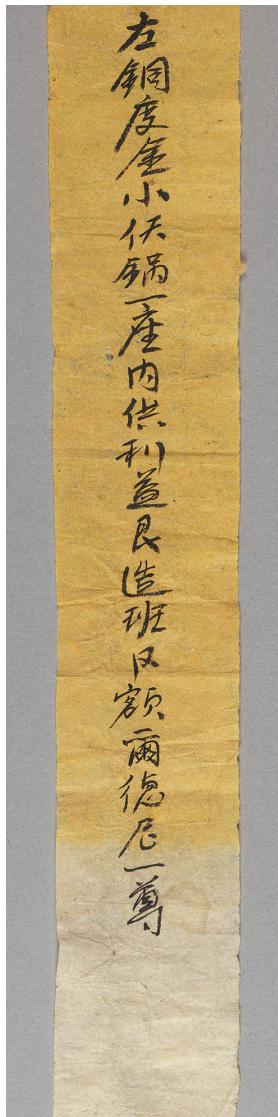
知识体系的缺失造成的信息组织层面的问题是多方面的，例如信息的重复和冗余、缺失链接数据价值而造成“数据孤岛”等，最为直观的挑战仍在“明明有却找不到”的检索利用效率方面。主要表现为互相交织的三点：第一，藏品名称信息过于富集，需要具备相当的领域知识才能有效切分和理解相关信息；第二，现有的藏品信息结构化程度低导致机器可读程度低，搜索系统难以对信息进行深度组织和管理；第三，概念间关系的缺位加剧了信息粒度和聚焦点差异导致的融合问题。

从传统知识组织的角度看，我国博物馆现有的藏品信息编目是遵循基于都柏林核心元数据元素集(Dublin Core Metadata Element Set，简称“都柏林核心”)等国际标准基础元数据字段，并按我国《博物馆藏品信息指标体系规范(试行)》《馆藏文物登录规范》(WW/T 0017-2013)等行业规范进行组织的，拥有其自身的知识组织体系(knowledge organization system，简称KOS)。然而，考察最无国际标准可循的中国古代艺术品的编目实践，上述三个利用侧面临的问题尤为突出，以故宫博物院三件院藏文物的命名为例：

1. 乾隆款掐丝珐琅三事—开光福寿万代纹夔耳方炉
2. 鲜于枢行草晚秋杂兴诗页
3. 红色暗花纱绣彩云蝠花卉金十团龙纹女蟒

上述名称中包含了款识、工艺、器类、纹饰、书体、作者、颜色、材质、功用等信息，按照具体类型的藏品命名规范次序进行排列，例如：珐琅器类为“(某某款)+细类+纹饰+器类”组合，如“景泰款掐丝珐琅缠枝莲纹梅瓶”，织绣类藏品一般格式为“颜色+(工艺)+纹饰+织法+(性别)+(单或棉或皮)+功用”组合，如“明黄

[图一]清宫旧藏文物中常见的
内务府黄签注记
故宫博物院藏



色彩云金龙纹妆花纱男夹朝袍”“青色缎钉绣珊瑚米珠寿双喜纹帽头”。此命名方式植根于传统，前有曹操高陵等汉魏大墓记录陪葬品名称的刻铭石牌，后有清宫旧藏中内务府所附黄签[图一]，得益于中文文字紧凑严整的结构与形式，藏品名称成为了基础信息最为富集的一个元数据字段。与西文相比，中文没有单词间的空格帮助划分词义单位，词汇中的成词更为密集，加之省略和连词、一字(词)多义等特性，本就造成此类藏品命名的切分有一定难度，编目时使用的专有名词和术语又高度依托于领域专家自身的知识体系，而这一体系并未内化至藏品信息资源之中，如上文中“三事”“开光”“页”“蟠”等词汇，即使正确切分出来，脱离了传统器物、书画、织绣类文物研究的知识背景，其词义仍难以准确理解和把握其传达的丰富信息。

除不利于缺乏领域知识背景的用户理解和获取信息外，这样的藏品名称虽对管理和研究而言严谨精炼，对计算机读取而言结构化程度低，无异于自由文本组合，文化遗产领域高度专业化的文本能够提供给机器学习的语料本就不多，加之很多词义还须依靠上下文推断，当前计算机在联想和推理上不及人脑，不利于编目信息的自动化处理。其潜在的危险还在于，由于藏品名称已经包含了足够多的信息[图二]，编目人员可能会忽略对对应的材质、形制、纹饰等元数据字段本身的完善和持续维护，进一步降低了藏品信息的结构化程度。非结构化的信息难以被计算机深度组织和有效管理，从中准确提取出标准化的关键词、元数据来建立索引，或对信息进行有效编码和向量化，以实现基于内容和语义的匹配，而只能依靠简单的字符串匹配，根据相关性进行排序的依据亦不足，无法实现高级匹配和准确排序。

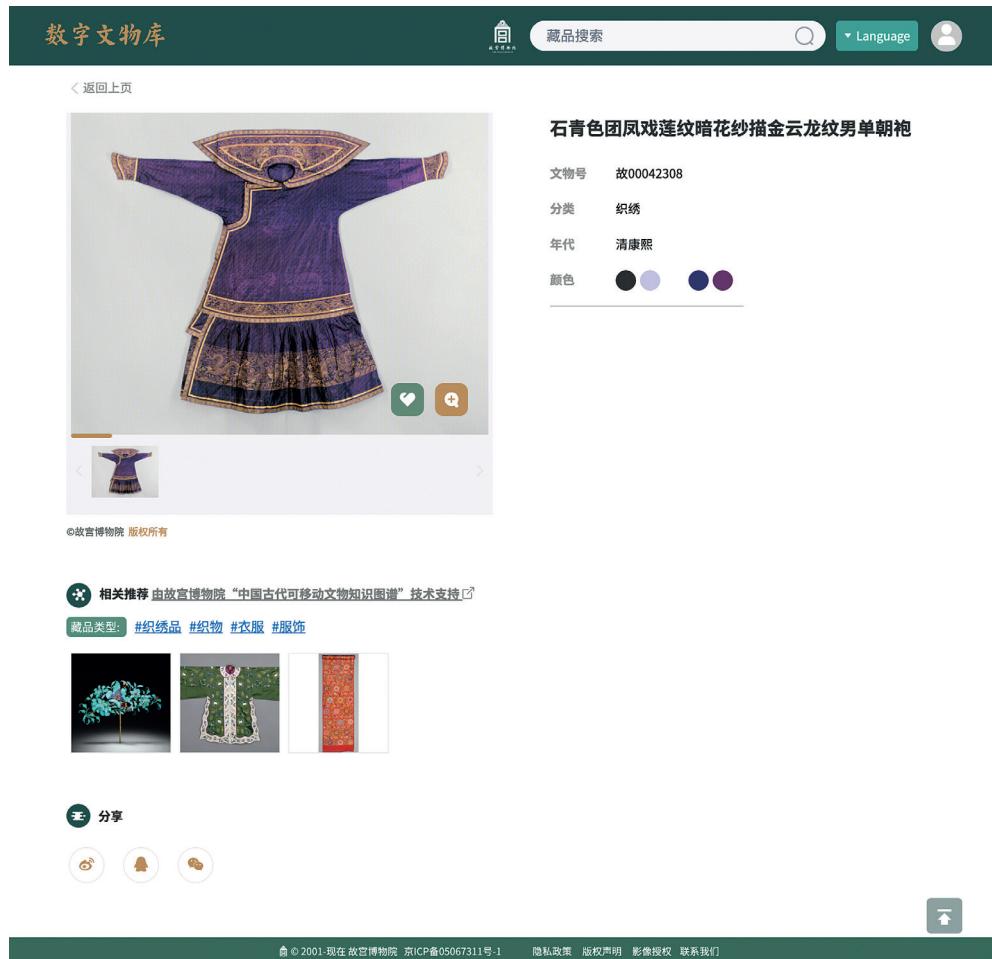
此外，编目中藏品的命名具有过程性和不确定性，对藏品的认识也是随着领域专家持续研究而由浅入深的。例如一件曾被断为烧制于隋代的青釉壶，可能因为参照新近考古发现的窑址和纪年材料而被重新认定为唐代器物；一件饰有“鱼龙纹”的器物，其装饰纹样在后续的研究中可能被细化为“摩羯纹”。此类情形容易造成藏品名称这一核心元数据的频繁更改和变动，且产生大量概念术语间关系问题。例如不同层级的概念术语并置——考察“永乐款剔红菊花纹圆盘”“宣德款戗金彩漆花卉纹圆盘”两个名称里对于纹饰的描述，前者使用的“菊花纹”的范围完全为后者使用的“花卉纹”范围所覆盖，后者是前者的上位词^①。当前常见的藏品信息组织手段仍为一套较为基础的受控词表(controlled vocabulary)，即用一致性的标签(在藏

^① 《GB/T 13190.1-2015/ISO 25964-1:2011, MOD信息与文献 叙词表及与其他词表的互操作》第1部分“用于信息检索的叙词表”，国家标准出版社，2015年，第2—5页。

品信息管理系统中表现为一系列的下拉菜单所提供的选项)来识别每个概念的有效性^①,包括同义关系、交叉关系、属种关系、整部关系等常见的概念间关系尚未有效确立。在靠字符串匹配进行检索的情况下,靠“花卉纹”无法查询到名称包含“菊花纹”的藏品。

由此可见,在大数据的时代,即使有了持续跃升的计算机算法算力加持,藏品知识组织的现状仍决定了对藏品信息资源的检索利用始终被局限于“句法”层面——藏品信息结构化程度低导致检索模式落后单一,而知识体系和概念间关系的缺失则导致用户理解困难、发起检索时未必能了解和使用那个“唯一正确的”关键词,当前阶段离基于机器可读的数据通过联想和推理所实现的内容和语义的精准匹配还十分遥远。检索效率低所反映的仅仅只是问题的一个局部,结构化和规范化程度低还会导致跨信息源的信息无法聚合检索。而分散建设的藏品信息资源将持续以孤立的数据集的形式存在,是无法有效支持应用数字人文的研究方法,通过计算机分析和挖掘来识别模式和规律,从而实现对藏品信息价值的进一步挖掘和洞察的。在一个用户普遍对谷歌、百度等头部搜索引擎的信息检索及智能推荐习以为常的时代,现有的博物馆藏品信息检索利用模式正在逐渐变得令管理者和终端用户都无法接受。

[图二] 故宫博物院藏品规范命名示例
这件朝袍的名称中包含了颜色、纹饰、材质、用途、物件类型等多方面信息



^① 受控词表是由表达概念的语词、标目或代码按规则构成的列表。通过受控词表进行的词汇控制,其目的在于增加标引人员和检索者选择相同语词表示特定概念的可能性。叙词表、主题词表和名称规范文档都是受控词表的实例。下文所提及的“艺术与建筑索引典”“中国图像志索引典”“文物主题词表”中“索引典”“主题词表”均为叙词表(thesaurus)的不同翻译形式,本文在叙述中遵循词表原编写者使用的表述方式。前揭《GB/T 13190.1-2015/ISO 25964-1:2011, MOD信息与文献叙词表及其他词表的互操作》,第3—10页。

二 “中国古代可移动文物概念参考模型”的构建：缘起与策略

与图书馆的知识组织颇为不同的是，博物馆的藏品更多时候是独一无二的历史遗存，对于个体藏品——尤其是那些具有重要历史价值、艺术价值藏品的知识组织，必须首先由其收藏机构打下基础：为其拍摄数字影像，提供标准化的元数据基础信息，将其相关的研究成果以最用户友好的方式组织起来。同时，由于艺术历史类博物馆藏品的时间跨度，来自同一位创作者、同一个遗址、同一历史时期、同一风格流派的藏品，常因流转而分散于世界各地的机构之中，为研究带来了诸多不便。互联网和大数据的时代，数字技术为对博物馆藏品知识的全局性组织提供了平台、工具，甚至是基础方法，然而真正的将这样的可能性变为现实，仍需从博物馆做起。正如2022年8月24日，国际博物馆协会（ICOM）在布拉格发布的博物馆新定义所言：“……博物馆促进多样性和可持续性……为教育、欣赏、深思和知识共享提供多重体验。”^①通过有效的知识组织服务于知识共享，是当下博物馆履行其使命的要义所在。

当下我国数字公共文化服务正处于由“有没有”“缺不缺”向“好不好”“精不精”转变的关键时期，对于如何更好地发挥博物馆数字资源建设的社会效益，响应日益增长的藏品信息查询利用的需求，面向中国古代艺术藏品的知识组织实践也进入了意识觉醒、体系选型和流程探索阶段。长期以来大量实践证明，在藏品编目中使用专业受控词表，将同义词加以整合，建立概念间的关联，是提升藏品信息查询和获取效率最基本、最有效的手段。当下国内文博界讨论藏品知识组织、数据开放获取等话题时备受推崇的Europeana^②、文化史图像数据库（Arkyves），及荷兰国立博物馆、纽约大都会艺术博物馆在内的文博机构开放高质量藏品数据集等各项最佳实践，均是在遵循包括“艺术与建筑索引典”（Art & Architecture Thesaurus，简称AAT）、图像志分类系统Iconclass在内的各种分类法、叙词表、权威档等标准化KOS来对藏品进行编目实现的〔图三〕，而这些体系无一不有着长达数十年、甚至超过半个世纪的漫长更新迭代、普及和权威化历程。由北京科技大学黄明玉教授牵头起草的文物保护行业标准草案“文物主题词表编制规则”以中国现行文物分类体系及AAT为基础，构建中国文物分类主题词表^③，英国维多利亚与艾伯特博物馆（Victoria & Albert Museum，简称“V&A博物馆”）的“中国图像志索引典”项目（Chinese Iconography Thesaurus，简称“中图典”）的数据库基于Iconclass顶层框架构建，面向海内外收藏机构中唐代至清代期间含有图像或纹饰的可移动艺术品、工艺品和古籍善本提供图像志受控词表，其在线查询系统仍在持续建设过程中。

① International Council of Museums. *Museum Definition*. <https://icom.museum/en/resources/standards-guidelines/museum-definition/>, 查阅时间：2023年8月1日。

② 刘芳《数字人文环境下博物馆藏品知识组织及应用的新思路》，《中国博物馆》2022年第3期，第75—80页。

③ 黄明玉《中国文物分类主题词表构建方法学刍议》，《图书馆学与资讯科学》2017年第1期，第171—190页。

(图三)Iconclass中对“重华宫茶宴连句”进行图像志标引的编号及上下文位置

The screenshot shows the Iconclass interface. At the top, there are links for 'ICONCLASS', 'Browse & Search', 'Documentation', and 'News'. On the right, there are buttons for 'Login', 'Feedback', and a forum icon. Below the header, there are two tabs: 'Browse' and 'Search'. The 'Search' tab is active, indicated by a blue background. The search query 'family of a ruler, and court' is entered. The results page displays a list of codes and their descriptions. The code '44B15 · family of a ruler, and court' is highlighted with a red border. To its left is a vertical list of related codes, and to its right is a detailed description of the term, including its broader context and specific sub-codes. A note at the bottom right says 'For more features, [you need to log in.](#)'

4 · Society, Civilization, Culture
44 · state; law; political life
44B · government, the State; 'Politica', 'Autorità o Potestà' (Ripa)
44B1 · ruler, sovereign

44B15 · family of a ruler, and court

Search with these related keywords:
Autorità o Potestà, Politica, Ripa, civilization, court, culture, family, government, occupations, royal family, ruler, society, state

Add more detail:

44B150 · 'Corte' (Ripa)
44B151 · family of the sovereign
44B152 · royal household; courtiers, retinue, train
44B153 · court ceremonial, court etiquette
44B154 · court festival
44B155 · daily life at court
44BB15 · family of a ruler, and court - BB - female ruler
44B15(+0) · family of a ruler, and court (+ variant)

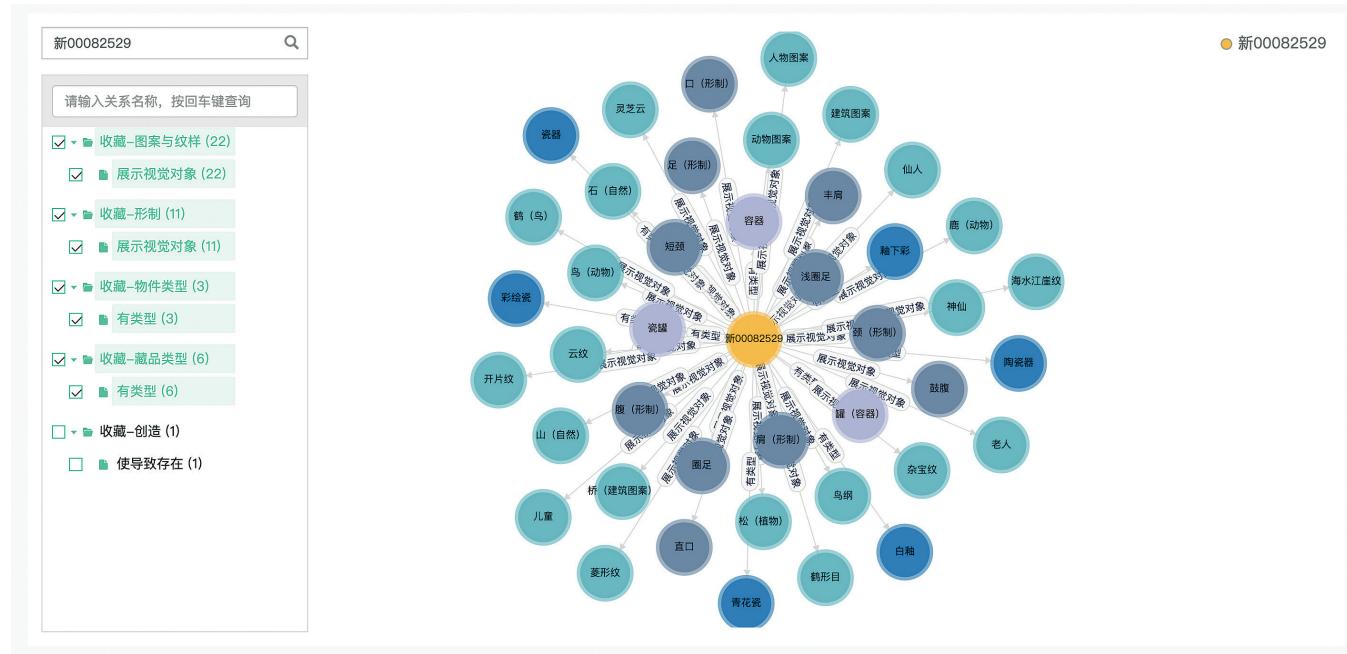
For more features, [you need to log in.](#)

结合中国古代艺术品为代表的我国博物馆藏品知识组织的现状，分析文化遗产行业在藏品知识组织方面最佳实践的案例，可以发现有效优化提升藏品信息检索和利用效率的解决方案，其本质是构建机器可读的规范化领域知识体系，并将其内化到机构基于编目建立的藏品信息资源数据集中，即在现有的博物馆信息化和数字化建设成果的基础上，采用机器可读的知识架构进行数据建设。在AAT、Iconclass等叙词表和权威档案代表的传统藏品知识组织体系基础上，大数据时代基于语义网络的知识图谱技术以结构化和统一的方式表达知识，一方面支持知识的直观组织和可视化交互，另一方面有助于更高效地构建机器可读的知识体系，从而推动藏品知识挖掘和利用的智能化。

2020年，故宫博物院获文化和旅游部科技教育司“文化艺术和旅游研究项目信息化发展专项”资助，开展“应用于人工智能搜索的可移动文物‘概念参考模型’研究”课题研究，利用两年时间，基于对故宫博物院院藏文物的知识组织，对国际标准、面向文化遗产领域的国际文献工作委员会概念参考模型(CIDOC CRM)进行拓展，构建“中国古代可移动文物概念参考模型”(Ancient Chinese Artifacts Conceptual Reference Model，简称CRM-ACA)，便是出于这样的考虑。CIDOC-CRM是文化遗产领域认可度最高的一款本体，它提供了一个顶层的领域模型框架，帮助博物馆、档案馆和图书馆将大量分散、不兼容的文化遗产信息基于共同的概念基础进行比较映射、数据转换等，表示不同数据格式中包含的共同的信息内容，从而实现信息的交换和集成，推动文化遗产信息的共享¹。经过二十多年的建设，CIDOC-CRM在2014年被纳入了

¹ Doerr, M. (2003). "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata". *AI Magazine*, 24(3), p.75. <https://doi.org/10.1609/aimag.v24i3.1720>, 查阅时间：2023年8月1日。

[图四]基于CRM-ACA在知识图谱平台丰富文物检索维度和概念间关系

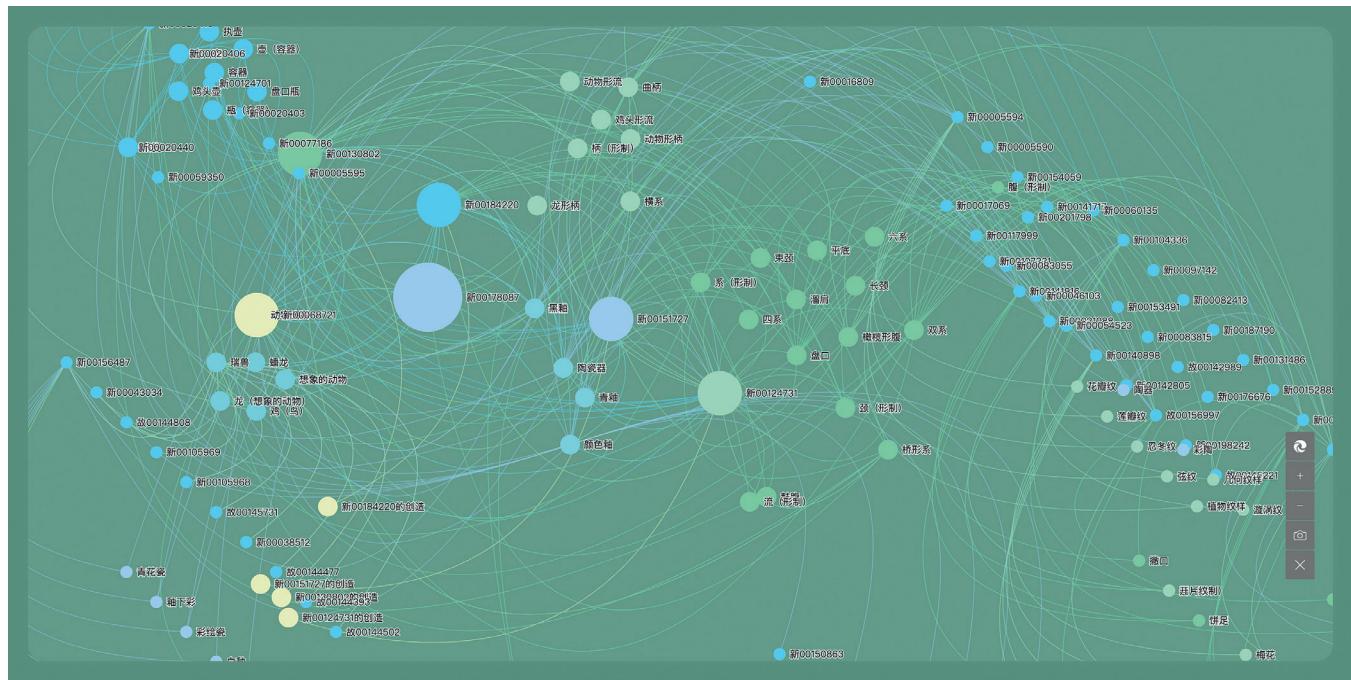


国际标准(ISO 21127:2014), 2020年被采纳为我国国家标准(GB/T 37965-2019 信息与文献 文化遗产交换的参考本体)。此外, 它所提供的高度精简、标准化的语义关系也为后续实现不同学科领域知识的交叉和融合打下了基础。

在国内文博机构中，故宫博物院180余万件(套)中国古代文化艺术品的收藏，体系完备，涵盖古今，品质精良，品类丰富，且拥有优秀的领域专家研究团队，成果卓著，在构建具有综合性的中国古代可移动文物领域本体方面有着独特的资源优势和专业优势。结合CRM-ACA建设的叙词表在纳入了“中国文物分类主题词表”“中图典”等主题性叙词表的基础上，参考藏品管理和研究部门配合数字化建设同步开展的藏品详细编目工作中对编目描述用语的进一步细化和规范，从物件类型、图案与纹样、藏品类型等不同分面构建概念间的语义关系。数十位来自绘画、法书、陶瓷、雕塑、宗教等多个领域的故宫博物院文物研究专家参与到了叙词表的建设与审核过程中，在确保叙词表的准确性和规范性的同时，也为相关词汇收录的范围和边界提供了领域研究和藏品管理业务场景的视角。

在藏品知识组织当中引入CRM-ACA，其目的并不在于颠覆领域专家现有的编目方式，而是基于现有的专家编目成果，利用语义网络技术，构建更丰富的规范化概念间关系（图四），赋予藏品信息资源以认知模型，从而支持计算机对藏品知识的深度理解。与此同时，通过数据标准化建设将这套本地数据集逐步处理成可以对齐全域的链接开放数据，纳入历史地理信息、人名规范等权威链接开放数据数据（LOD）对藏品数据集进行“富化”，从而让故宫博物院的藏品数据集，首先提升检索利用效率，逐步支撑规模化分析挖掘及更多的智能化应用场景，成为未来文化遗产领域数字人文基础设施中有价值的局部。此外，故宫博物院中国古代艺术藏品体系完备的价值在于，CRM-ACA的构建，及其对“中国文物主

[图五]“数字文物库”中“中国古代可移动文物知识图谱”可视化探索模式
在该模式下，用户可直观地探索经过了数据标注的文物间存在的概念关联



题词表”相关部分和“中图典”等受控词表的对接与纳入，是怀着以自身藏品知识组织体系建设为引领与示范、为中国古代可移动文物领域工程级数字人文基础设施奠基的雄心的。拥有中国古代可移动文物收藏的文博机构，可直接使用CRM-ACA进行藏品知识组织，或根据自身藏品特征对模型进行细化拓展，从开展藏品知识组织实践之初保障藏品数据的标准化、体系化，减少后续全域数据对齐可能导致的返工与重复劳动，以期未来不同机构的藏品数据建设成果共同助力中国古代可移动文物的研究与价值挖掘。

三 人工智能赋能文化遗产研究：展望与意义

2023年2月，基于CRM-ACA建设的“中国古代可移动文物知识图谱”正式应用于故宫博物院藏品信息资源公共文化服务平台“数字文物库”[图五]，通过持续的数据建设，为“数字文物库”中的藏品提供语义检索辅助和知识图谱的可视化探索。同期，以ChatGPT为代表的大语言模型(LLM)的成熟引发了国内外的高度关注，大规模无标注数据训练的模式从表面上看似乎颠覆了自上而下构建领域模型，并通过人工标注数据进行数据建设的模式。然而，大语言模型的成熟与领域模型的构建有着相辅相成的关系，前者为后者提供了广泛的语言知识和强大的预训练基础，而专精的领域模型则为前者提供特定领域的语言特征和知识需求，辅助实现该领域高准确性、高性能的语言理解与生成。若能拥有高质量的中国古代可移动文物领域模型，配合藏品数据建设，有了LLM的加持，人工智能技术的应用似乎更加触手可及。

耶鲁大学教授乔治·库布勒(George Kubler)在其著作《时间的形状》中写道,“造物的历史意图在视觉

形式的量规之下实现思想与对象的重聚：‘视觉形式’这一术语包括手工艺品、艺术作品的唯一实例及复制品……时间的形状即现于所有这些造物之中”^①，即手工艺品和艺术品呈现了创作者对时间的独特理解。以故宫博物院为代表的中国古代艺术博物馆的藏品正是通过研究展现过往时间之形的基石所在，正因如此，这些藏品的知识组织与开放共享方才尤为重要。与西方博物馆相比，国内文博机构藏品知识组织的开展仍然非常缺乏系统的方法论及数字化基础设施的支撑，甚至对于很多机构而言，藏品的基础信息化建设至今仍未能有效开展。因此，面向人工智能和数字人文研究的藏品知识组织实践往往要与前一阶段的藏品的信息化、数字化建设同步开展，我国博物馆在“补课”的同时仍须负重向前奔跑，方能真正为文化遗产领域的数字人文研究专家学者们准备好材料。CRM-ACA作为一种对可移动文物为物证的人类社会历史进行描述的知识结构模型，结合其在故宫博物院藏品知识组织当中的应用示范，希望能够为这一努力提供一种方法、一套工具，以期与有志于此的文博及学术界同仁，协力为我国文化遗产数字人文基础设施建设添砖加瓦，开启人工智能赋能的文化遗产研究新篇章。

[作者单位：故宫博物院数字与信息部]

(责任编辑：张 露)

^① Kubier, George. *The Shape of Time: Remarks on the History of Things*. New Haven: Yale UP (1962), p.9.