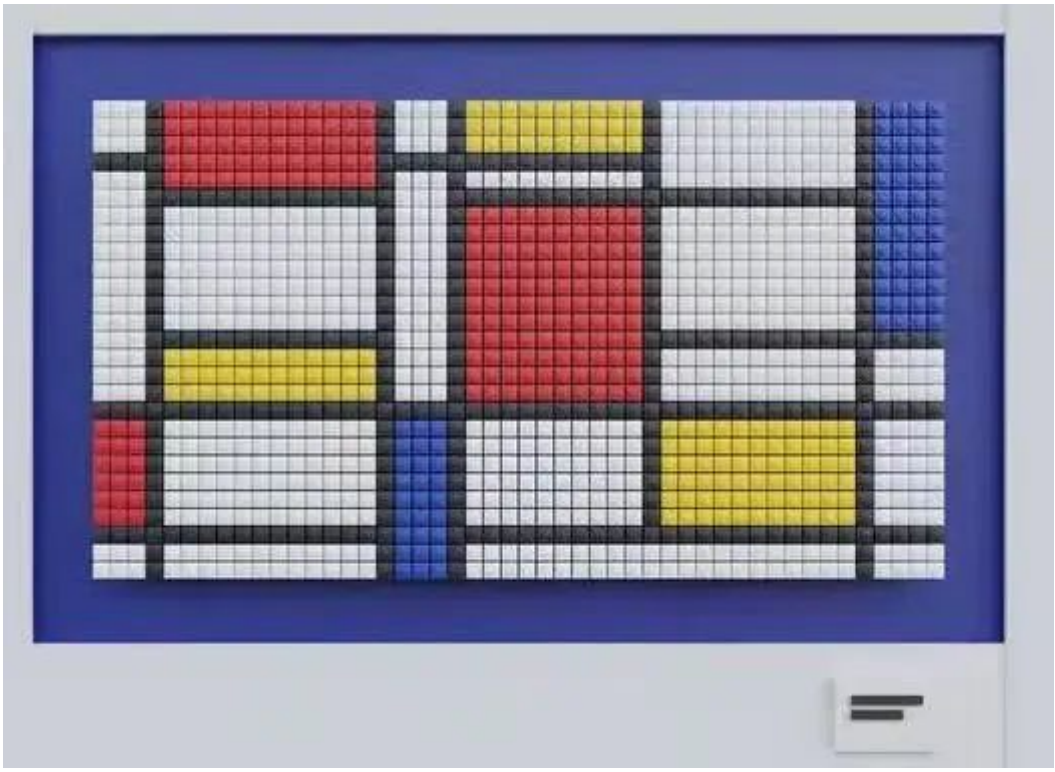

Photo by SIMON LEE on Unsplash.

夏翠娟
(上海图书馆，上海，200031)

作者按

在刚刚落幕的CDH2021的获奖论文中，评分靠前的三篇①均与“本体（Ontology）关，涉及多个不同领域。这意味着本体在数字人文的大帐篷下具备了多学科视野并进入应用实践阶段。本体发端于语义网（Semantic Web），GLAM领域作为排头兵，积极响应、稳步发展、深入应用，已形成了较为系统的方法论和标准规范体系。知识图谱的勃兴更是让本体在计算机科学领域得到了IT人员的亲睐。现在各个领域的人文研究者也都开始用本体实现多维、多面、多粒度的数据建模——领域专家来构建领域知识本体，可以说是用得其所。

笔者在前辈的引领下，入行已近20年，一直勉力跟踪国内外图档博和语义网领域关于本体的进展，也设计过多种历史文献资源、世界知识的本体模型、词表和编码规范，也曾和历史、艺术领域的学者合作过，接受本行业外的检验。然而，在AI技术高速发展的今天，免不了心中存疑，这种人为的本体设计是否还有用武之地？疫情期间，肉身被禁锢，恰是反思时，正好把这些年来设计本体的些微经验和浅薄思考，结合国际上的新进展和新趋势，草成一文，幸得《图书情报知识》垂青，发在2021年第1

期。希望有机会能与人文领域的研究者分享一个图书馆员关于本体设计的思路，做一个抛砖者，期待与更多的领域专家交流，得到人文学者的批评指正。

①CDH2021前三名获奖论文分别为：1.全生命周期下的明清官式建筑遗产知识的本体建模（天津大学：沈孙乐 马昭仪 何捷）；2.非物质文化遗产传统技艺知识本体构建——以锦、绣、年画三类为例（南京大学艺术学院、上海图书馆：李梦琦 夏翠娟 陈静）3.基于 OWL-Time 的时间本体在古籍中的应用（北京大学信息管理系、北京大学数字人文研究中心：王林旭 位通 王军）

目的/意义

探索美术馆、图书馆、档案馆、博物馆等机构（GLAM）中多源异构文化记忆资源跨领域跨机构的知识融通解决方法和路径。

研究设计/方法

在“元数据应用纲要”的基础上，提出了“本体应用纲要”的概念，以上海图书馆手稿档案元数据应用纲要扩展为本体应用纲要为例，阐述了本体应用纲要设计的原则、方法和流程；以历史人文“数据基础设施”建设为目的进一步构建了一体化本体及其知识融通模型，并验证基于本体应用纲要在一体化本体框架下的扩展以实现跨领域知识融通的方法。

结论/发现

元数据应用纲要为文化记忆资源的知识融通提供了结构化数据储备，本体应用纲要兼顾了特定资源类型的个性化需求，一体化本体的知识融通模型则用于保证知识建模的统一性和知识表示的一致性。

创新/价值

基于多种本体应用纲要的一体化本体设计方法和关联数据与知识图谱技术结合后，作为一种新的知识组织方法，为跨机构跨领域的知识融通提供了方法和路径。

关键词

本体应用纲要 知识融通 文化记忆

数字人文 数据基础设施

1 引言

美国著名社会生物学奠基人爱德华在《知识大融通：21世纪的科学与人文》一书中说：“科学和人文是由同一台纺织机编织出来的”。他认为，无论是自然科学领域的物理、生物、信息科技，还是人文学科的历史、文学、艺术领域，在知识的层面，必然存在着某种统一的秩序和广泛的联系，而建立这种秩序和联系的过程，称之为“知识大融通”[1]。对于自然科学来说，构建统一的知识体系的可能性比人文学科大得多。而人文学科被划分为许多独立的门类，尽管强调专业上的用词必须准确，但是不同的专业很少采用相同的术语。随着数字人文的勃兴，数据驱动的量化计算、可重复验证的实验、基

于高层次的抽象理念和原理创造的新工具等自然科学领域的研究范式被引入人文研究中。在人文学界，跨专业的交流也变得愈加频繁，基于互联网的数据传输交换和基于机器智能的跨学科知识整合创新创造了新的知识生产模式和交流环境。“知识大融通”不仅是一种梦想，而且是亟待解决的需求。

“知识融通（Knowlege Consilience）”与“知识组织（Knowledge Organization）”和“知识聚合（Knowledge Aggregation）”既有联系，又有区别。“知识融通”建立在“知识组织”对知识的整理、加工、揭示、控制和有序化组织的基础上，涵盖“知识聚合”对知识单元进行凝聚、以形成多维多层且相互关联的知识体系的过程[2]，但更强调跨学科跨领域的知识互操作，通过自上而下的知识组织来抽象出独立于领域、与资源的载体和格式无关的高层互操作层，通过自下而上的知识计算来构建知识间统一的广泛的联系。知识组织、知识聚合和知识计算是达到知识融通的方法和手段。具备跨学科特性的数字人文对“知识融通”的需求为GLAM（美术馆、图书馆、档案馆、博物馆等文化记忆机构的简称）领域提供了用武之地。这些文化记忆机构是文化记忆资源的集散地，长期致力于资源采集加工、知识组织与传播传承、跨机构资源共建与数据共享，为文化记忆资源的知识融通奠定了一定的知识储备与方法论基础。

在互联网、大数据和机器智能时代，“知识融通”往往发生在机器世界中，由大规模、多种类、跨领域、粒度不一、格式各异、来源不同的数据驱动，通过定量到定性的分析达成。机器世界中的“知识融通”要求统一的知识建模和一致的知识表示，前者作为高层、抽象的知识共享模型，超越学科话语体系，在不同领域的术语间建立映射和关联；后者提供一致的知识编码规范，以支持与原生数据格式无关的数据传输、交换和融合。GLAM领域基于分类主题法、元数据和本体的知识组织方法，结合关联数据、知识图谱和机器学习技术，可为机器世界的“知识大融通”提供一定的路径和解决方案。GLAM领域的元数据方案（包括元数据描述规范和编码格式）由于所描述的资源对象不同而千差万别，而本体则提供了一种抽象的高层视角，用面向对象的方法，通过对客观现实世界的模拟、对主观知识的客观表达、对隐性知识的显性表示，将异构资源元数据方案纳入到统一的知识模型中，用一致性的知识表示格式编码后形成本体应用纲要（Ontology Application Profile），实现机器世界中跨领域的知识融通。本文将以上海图书馆历史人文“数据基础设施”中的本体设计为例，探讨如何利用基于本体的知识组织方法，将众多遗留系统（Legacy Systems）的不同元数据方案和元数据应用纲要（Metadata Application Profile）升级（upgrade）为本体应用纲要，基于一体化的本体模型进行知识建模和知识表示，助力跨领域的知识融通。

2 现状调研

分类主题法、元数据方法、本体方法代表了GLAM领域知识组织方法发展的三个阶段，在前两个阶段，GLAM分别有各自的领域标准，而随着本体方法的广泛应用，逐渐呈现出跨领域融合的趋势。

在图书馆界，试图改善MARC的封闭性、复杂性和专业性的DC元数据及其方法论得到广泛而深入的应用，超越了图书馆的范围，延伸到GLAM领域和互联网，在国内成为科技部、国家图书馆推出的涵盖纸本资源、数字资源、网络资源、多媒体资源的数字图书馆系列标准规范的基础。对图书馆领域产生深远影响的国际图联FRBR模型及其家族FRAD、FRSAD为基于本体的知识组织方法奠定了基础。它以“实体-关系”分析方法来重新构建书目记录、规范记录、主题规范记录的功能需求框架，于2017年整合升级为“图书馆参考模型（IFLALRM）”，旨在更好地支持图书馆的关联数据应用[3]。图书馆界广受瞩目的关联书目数据模型BIBFRAME是旨在将MARC数据转换为关联数据的本体模型和词表，它的三层核心模型（作品-实例-单件）是FRBR四层模型（作品-内容表达-载体表现-单件）的简化。除了核心模型外，BIBFRAME还包括人物、机构、地点、事件、注释等类和属性，设计初衷不仅仅是应用于图书馆，而是希望能描述所有的GLAM资源[4]，其创始人之一EricMiller领衔的Zepheira团队维护了数个BIBFRAME的本体应用纲要，其中包括用于档案资源描述的BIBFRAMELite+Archive②。BIBFRAME于2016年发布了第二版，目前正由美国国会图书馆领衔的各类图书馆测试和完善。本体应用纲要使得在统一的BIBFRAME本体模型框架下，仍然能兼顾不同领域的知识组织需求，同时促进跨领域的知识融通。

②http://bibfra.me/view/archive/

在博物馆界，用于描述艺术作品或作品集合、建筑、其它物质文化遗产及其数字化影像的本体包括盖蒂中心（TheGettyCenter）维护的CDWA、国际文献工作委员会（CIDOC）的CIDOC-CRM。最新版的CDWA于2019年9月更新，包含分类、名称、创建者、物理形态、材料、标记、语境、时间、地点、所有者、流藏历史、相关作品、主题规范等31个大类，共540个类目（Category）。CDWA的“类目”可理解为元数据项，只有解释性定义，没有定义域和取值范围的说明，所以它并不是数据模型，但可以为设计数据模型提供概念参考框架。CIDOC-CRM的第一版于国际图联推出FRBR后不久发布，试图在博物馆的各种文物信息之上建立一个抽象的概念层[5]，以支持文物信息的整合、交换、共享和重用。与FRBR相比，CIDOC-CRM更进一步地采用了面向对象的思想[6]，明确地定义了文化遗产领域的各种实体（概念）、属性（关系），并发布成可共享重用的术语词表[7]。CIDOC-CRM在博物馆领域得到了广泛的应用[8]，早在2006年，就有学者和工程师研究利用该模型和语义网技术构建多博物馆文物信息集成的虚拟博物馆[9]；经过多次的扩展，除了文物本身外，CIDOCCRM能够描述与文物相关的历史事件、艺术特征、考古遗迹、时间、地点、人物等信息，于2014年成为文化遗产领域的国际标准（ISO21127:2014）。此外，CIDOC-CRM得到了图书馆领域的关注，为了支持与FRBR之间的

互操作，2007年成立了专门工作组来维护CIDOC-CRM与FRBR之间的映射，FRBR也开始借鉴CIDOC-CRM的面向对象分析法，形成FRBRoo框架[10]；近年来，基于CIDOC-CRM本体模型和词表的知识组织方法也被应用到数字人文实践中[11,12]。

与图书馆界相比，档案界的元数据标准制定稍显滞后，国际档案著录标准（ISAD）由国际档案理事会（ICA）于1990年开始制订，与图书馆MARC齐名的档案元数据标准EAD的第一版于1998年完成。EAD基于ISAD设计，著录详尽且适用范围广泛，其比MARC更有先天优势的地方在于一开始就应用了XMLDTD来定义元数据元素，具有较好的易用性、兼容性和可扩展性。与图书馆界的元数据标准不同的是，档案的描述还要强调档案产生的背景，因而还有ISAAR(CPF)和EAC这样用于描述与档案有关的团体机构法人、个人及其家庭背景的元数据规范和编码规范，可用于GLAM领域的个人传记、机构沿革、家庭历史信息的编码。其中ISAAR(CPF)是ISAD的补充，在欧洲国家应用广泛，EAC是EAD的延伸和扩展，在美国应用较多[13]。尽管国内有基于CIDOCCRM来构建数字档案资源描述框架的研究[14]，国外有复用多种GLAM领域已有模型和词表来建立领域本体词表的案例，例如西班牙国内战争图像档案项目[15]，但是档案领域尚未产生具有深刻影响的本体模型。

从国际范围来看，本体方法促进了GLAM领域的知识融通，尤其是“应用纲要”的应用，在兼顾领域特殊性的同时保证知识模型的统一性与知识表示的一致性方面提供了方法和路径。2019年的DCMI国际年会举行了一个以“应用纲要”为主题的研讨会，旨在探索应用纲要的方法在不同领域中的应用。

3 从元数据应用纲要到本体应用纲要

3.1

元数据与本体的联系与区别

元数据（Metadata）的定义是“关于数据的数据（Data about data）”，本体（Ontology）特指W3C语义网技术框架中的本体，非哲学意义上的本体，其定义是“对概念的明确的、形式化、可共享的规范说明”[16]。元数据和本体在本质上都是信息的结构化描述方法，但在构成、描述对象、描述重点、词表定义、标准规范、编码、存储与查询技术方面存在着一定的区别，详见表1。

表 1　元数据与本体的区别与联系

		元数据	知识本体
相同点		对信息的结构化描述	
不同点	定义	“关于数据的数据。”	“对概念的明确的、形式化、可共享的规范说明。”
	构成	元数据元素、取值词表，元数据记录	模型、术语词表（类、数据属性、对象属性），形式化编码，实体/实例对象
	描述对象	文献、信息资源	知识
	描述重点	文献资源的特征	概念、概念的特征及其关系，领域知识的规则和公理
	词表定义	元素、元素修饰词、编码体系修饰词	类、父类，属性、父属性、域、范围，规则，公理
	标准规范	MARC/EAD/DC	FRBR/EDA/CIDOC-CRM/BIBFRAME/FOAF/schema.org
	编码	ISO2709/RDF/XML	RDFs/OWL/SKOS
	存储与查询	RDB/SQL	RDF Store/Graph DB/SPARQL
联系		元数据元素可作为本体中概念的属性；本体可看做是关于元数据的元数据，为不同元数据方案提供高层互操作的解决方案。	

元数据标准规范一般由元素集（也叫“元数据术语词表”，在MARC中是字段和子字段）、著录规范、编码规范组成。特定应用领域的元数据方案是对元数据标准规范在特定领域应用的进一步规定，除了采用一定的元数据标准规范外，还会在标准规范的元素集基础上进行扩展，以形成适用于特定应用领域的扩展集，并制订更为细致的著录规则和与系统实施技术相关的编码规则。元素集是元数据标准规范和元数据方案的核心组成部分，一般由元数据元素（element）、元素修饰词（refinement）或称子元素（subelement）、编码体系修饰词（Scheme，元素的取值词表，一般来自于某个规范的受控词表）组成。

一个完整的本体一般应包含三个层面：①模型（概念和概念间的关系）、②术语词表（以特定的明确的语词表征的概念、概念特征、概念和概念间的关系）、③机器可读的形式化编码。术语词表由“类（Class）”和“属性（Property）”组成，其中用来表征概念的术语叫“类”，表征概念特征和概念间关系的叫“属性”。对本体设计者来说，往往关注的是模型和术语词表，其形式化编码容易被忽略，但却是语义网框架下本体实施的关键，因为本体是为机器服务，目的是促进机器之间的语义互操作。对术语词表进行形式化的本体语言有RDF Schema(RDFs)，OWL，SKOS，其中OWL和SKOS是RDFs的子集。RDFs可以定义类及其子类、属性的范围和域。与RDFs相比，OWL具有更强的定义能力：根据属性的取值为数据还是为另一个对象将属性区分数据属性（dataProperty）和对象属性（objectProperty）；定义术语之间的关系，如等同关系owl:equivalentClass、owl:equivalentProperty，同指关系owl:sameAs等；还可以定义规则和公理，用于知识挖掘和推理。SKOS主要用于对知识组织体系中的概念词表如索引典（thesauri）、分类法（classification schemes）、主题标目（subject heading lists）、分类体系（taxonomies）、大众分类法（folksonomies）等受控词表进行形式化定义，可定义概念的各种标签（skos:altLabel）、象征

（skos:altLabel）、概念间的关系，例如一个概念的上位概念（skos:broader）、下位概念（skos:narrower）、顶层概念（skos:hasTopConcept）等。SKOS的高级扩展还可以定义不同概念词表之间的映射。

元数据以文献、信息资源为描述对象，描述重点在于资源的物理特征（例如对印本书的题名、著者、版本、载体形态的揭示，对电子资源的格式和存储获取路径的揭示），目的是管理、检索和获取。本体以知识为描述对象，粒度不拘、种类多样、范围广泛，可以描述文献、信息资源，还可以描述其内容中隐含的知识，如人、机构、地点、时间、事件、物体、主题词、关键词等实体或概念的特征及其相互之间的关系。元数据是以资源为中心的辐射结构，本体则是去中心化的立体网状结构，任何描述对象都可以作为管理、检索、获取的切入点。元数据元素可以作为本体中概念的属性，例如“题名（dc:title）”这个元数据元素可以作为“音乐（shl:Music）”这个类的属性，而本体则可以看作是关于元数据的元数据，是元数据的进一步抽象和规范，为不同的元数据方案提供高层的互操作方案。

3.2

元数据应用纲要和本体应用纲要

应用纲要（Application Profile）由都柏林核心元数据组织（DCMI）于2007年在新加坡举行的DC元数据国际年会上提出，被称为“新加坡框架（Singapore Framework）”或“DCAP（Dublin Core Application Profile）”。该框架为DC元数据标准规范在特定领域中的应用提供了理论框架和实施流程上的指南，为面向实际应用的元数据方案的组成和实施步骤提供参考规范。在“新加坡框架”中，应用纲要的组成部分包含：经过功能需求分析后建立在领域标准基础上的领域模型，使用已有元数据标准规范词表的、利用RDF（s）语言且依据DC抽象模型（DCAM）进行形式化定义的元素集和建立在DCMI句法指南规范基础上的编码规范与数据格式规范三个部分[17]。可以看出，应用纲要是对已有模型和词表在特定场景应用实施的进一步细化，并强调编码和句法，也就是说应用纲要是面向机器、以规范的形式化语言固化下来的元数据应用规范，便于在相同领域和相似场景中共享和重用，以促进不同元数据方案之间的语义互操作。自“新加坡框架”提出之后，应用纲要成为一种系统性的方法论，真正将DC元数据从标准规范的层面推向各个具体的应用领域，在整个GLAM世界得到广泛深入的应用[18]。

应用纲要的方法也在互联网尤其是语义网领域得到了继承和发扬，W3C将其定义为特定应用领域的数据描述规范，包括从已有的一个或多个标准中复用术语，同时根据具体应用需求更进一步定义术语的必备性和可重复性、取值范围或推荐使用的受控词表。根据KarenCoyle的定义，“应用纲要”是为特定应用设计、描述数据集的内容和结构、机器可操作、与应用程序无关的文档；至少包括三个部分：

描述数据集的数据元素、元素取值约束、元素的人读标签与说明。应用纲要应同时满足人读和机读的需要，可用于数据的验证[19]。

BIBFRAME（书目框架）作为旨在取代MARC的关联书目数据框架，也采用了应用纲要的方法，是应用纲要方法在语义网技术驱动下GLAM领域应用的典型范例。BIBFRAME项目制订了书目框架应用纲要(BIBFRAMEProfile)规范[20]，是如何将BIBFRAME核心模型和本体词表应用于具体领域的指南性规范，定义了如何为领域应用构造一个应用纲要的规则和语法，以支持BIBFRAME的核心本体模型和词表在特定领域的应用实施。书目框架应用纲要具体表现为一个或多个文件，以一定的格式编写而成，且可被机器处理。它由遵循一定语法规则的“纲要定义(ProfileDefinition)”“资源模板(Resources Templates)”“属性模板(Properties Template)”三个组件构成。“纲要定义”声明该应用纲要描述的领域资源类型，“资源模板”规定具体应用纲要包含的类，“属性模板”定义一个类包含哪些属性、各个属性的域与范围以及属性值的数据类型约束与取值约束。上海图书馆基于BIBFRAME设计的家谱本体即是一个书目框架应用纲要的例子[21]。

笔者为了规范领域本体设计的方法和流程，既保证不同资源类型本体在模型上的一致性和词表的可重用性与一致性，同时又充分满足描述与揭示的特殊性和多样性需求，故提出“本体应用纲要（Ontology Application Profile）”的概念和基于“应用纲要”方法的本体设计方法和流程。本体应用纲要是一种特别的“应用纲要”，沿用“应用纲要”的定义：是为特定应用设计、描述数据集内容与结构、机器可操作且与应用程序无关的文档。相对于元数据应用纲要本体应用纲要的组成部分有所不同，根据本体的特点，至少包括三个部分：表示领域概念及其关系的抽象模型，由特定机构或组织维护、有着特定命名空间的术语词表，以及同时支持人读和机读的对每个术语的形式化定义。本体应用纲要的设计遵循以下原则：①模型尽可能基于现有领域本体模型扩展，以最大限度地支持领域模型的共享和互操作；②术语尽可能复用现有的本体词表，复用术语时不改变该术语维护机构给出的定义，也可依据应用需求自定义新的术语；③依据特定应用需求以机器可读的形式化语言（JSON, RDF, XML）定义术语的应用规则和约束，对复用的术语在与原定义不冲突的基础上，可进一步定义其在特定应用中的各种应用规则和约束，同一术语在不同的应用纲要中的应用规则和约束可以不同。

3.3

从元数据应用纲要到本体应用纲要设计

为了说明一个特定领域的本体应用纲要在设计过程中如何利用已有的元数据应用纲要的成果，如何遵循上述三个原则，笔者以名人手稿的本体应用纲要为例进行阐述。

上海图书馆名人手稿元数据应用纲要中的元数据元素集分为核心集和扩展集，复用了DC元数据规范中的15个核心元素和DCTerms的部分元素，同时根据名人手稿12种不同资源的特殊需求自定义了若干元素[22]。元素集的定义主要是对元素及元素修饰词的定义与取值约束的声明，如图1所示。元素尽可能地复用已有的在Web上公开发布的元数据术语词表，但在应用纲要中可进一步对元素和元素的取值进行具体的约束，如对元素的“必备性”和“可重复性”的规定，对元素取值的数据类型约束或取值词表的约束，即指定取值须来自某个标准规范的词表，称为编码体系修饰词（Scheme）。以复用自DCTerms词表中的dct:created元素为例，对元素和元素值的各项约束如表2所示，并以XMLSchema进行形式化编码。

本体应用纲要的组成和定义框架如图2所示，包括类与属性的定义及属性与属性值的约束，其中类和属性尽可能地复用已有的术语词表。上海图书馆名人手稿本体应用纲要在元数据应用纲要的基础上设计。基于以BIBFRAME2.0为核心模型的家谱本体应用纲要扩展而来，元数据应用纲要中元素描述的对象以手稿档案资源为中心，在本体应用纲要中，手稿档案资源按照BIBFRAME的三层模型分为作品（bf：Work）、实例（bf:Instance）、单件（bf:Item）三个层面，除此之外还定义了人（shl:Person）、机构（shl:Organization）、地点（shl:Place）、时间（shl:Temporal）、事件（shl:Event）等与资源相关的类。部分描述手稿资源的属性复用了元数据应用纲要中的元素。但本体属性的定义与元数据元素的定义相比，除了对“必备性”和“可重复性”的约束外，还增加了域（Domain）和范围（Range）的约束，域指定属性所描述的对象，范围即是对属性取值的约束。与元数据元素相比，属性的取值还可以是另一个类，使得属性的定义具备了更为丰富和明确的语义。表3以自定义的shl:Resource和复用自BIBFRAME2.0的bf:heldBy属性为例，说明了在本体应用纲要中如何通过对已有类的继承（subClassOf）定义新的类，如何在复用属性原始定义的“域（Domain）”和“范围（Range）”之外，根据领域应用的特殊需求对属性的“必备性（Mandatory）”“可重复性（Repeatable）”“用于描述（UsedWith）”的类和“预期取值（ExpectedValue）”进行更为具体的定义和约束。在本体应用纲要中，“用于描述（UsedWith）”的类应和原始定义的“域（Domain）”保持一致或为其子类，“预期取值（ExpectedValue）”的类与原始定义的范围（Range）保持一致或为其子类，这样就不会与原始定义产生逻辑上的冲突，笔者称之为“术语复用一致性与应用纲要差异性原则”。如对bf:heldBy属性规定其“用于描述（UsedWith）”的类为“bf:Item”，与原始定义保持一致，规定其“预期取值（ExpectedValue）”的类为shl:Person和shl:Organization，是原始定义“范围（Range）”的bf:Agent的子类。

4 异构资源本体应用纲要与一体化本体设计

4.1

异构资源本体应用纲要设计

家谱、手稿档案、古籍以及与上海历史文化相关的各种电影、音乐、老照片、近现代期刊报纸、上海优秀历史建筑、物质文化遗产名录等（下文统称“上海记忆”）都是GLAM领域典型的特色资源，由于各种原因，其著录所用的元数据标准规范和格式不尽相同。其中7万余种家谱以《中国家谱总目》的联合目录数据为基础，包括上海图书馆的馆藏，源于597家包括图书馆、档案馆、博物馆、纪念馆在内的机构家谱馆藏，以及若干宗亲会和私人收藏，数据格式为半结构化文本和MARC格式。7万余种名人手稿档案来自上海图书馆文化名人手稿馆多年来的收藏，元数据记录的格式为DC/XML。100万余条古籍书目数据来源于上海图书馆的馆藏古籍目录，以及各种现代联合目录、历代史志目录、官修目录、藏书楼目录、私家藏书目录、避讳字知识和印谱等，格式主要为MARC格式和半结构化的文本。“上海记忆”的资源主要来源于上海图书馆的馆藏和文旅机构发布的数据，多种类、多媒体、多格式是其显著特征。上海图书馆近年来利用数字人文理念、方法和技术试图建设一个兼容各种GLAM机构资源类型的历史人文大数据平台，为数字人文研究提供“数据基础设施”。其中本体作为一种与关联数据、知识图谱技术相结合的知识组织方法，为异构资源的描述和揭示提供了自上而下的知识建模和自下而上的知识表示路径，通过本体应用纲要中术语的复用和继承可以实现跨GLAM领域的知识融通。

于2014年开始设计的家谱本体应用纲要复用了2014年发布的BIBFRAME模型和词表，于2016年之前升级到BIBFRAME2.0，除了以BIBFRAME的词表来描述家谱文献的各种信息之外，还自定义了人（shl:Person）、姓氏（shl:FamilyName）、机构（shl:Organization）、地点（shl:Place）、时间（shl:Temporal）、事件（shl:Event）及其子类迁徙事件（shl:MigrationEvent），并定义了各种类特有的属性，如shl:Person继承了foaf:Person类和bf:Person类，还定义了各种中国人独有的属性，如谱名、字、号、字辈、排行等。紧随其后设计的手稿档案本体应用纲要复用了家谱本体模型和词表，并自定了印章（shl:Seal）、藏书票（shl:BookPlate）、信封（shl:Evnvelope）、信纸（shl:LetterPaper）、多媒体特征（shl:MediaCharacteristic）等手稿档案特有的类及其属性。于2017年设计的古籍本体应用纲要，在已有的家谱和手稿档案本体应用纲要的基础上，复用了其中自定义的人、地点、时间、事件等类和属性，但由于古籍文献特征的丰富性和作品、版本、单件之间关系复杂性，无法完全复用BIBFRAME2.0模型，而是继承了其“作品-实例-单件”的三层模型并加以扩展，自定义了作品（pmb:Work）、版本（pmb:Instance）、单件（pmb:Item）、分类体系（pmb:Classification）、注释（pmb:Annotation）这五个核心类及其属性。2018年设计的上海记

忆本体应用纲要在家谱、手稿档案、古籍本体应用纲要的基础上进行大幅扩展，继承和复用了DBPediaOntology,schema.org的本体模型和词表，资源类型除了图书类型的文献外，还增加了期刊报纸类型的连续出版物文献，另外还有电影、音乐、老照片等多媒体资源以及优秀历史建筑和物质文化遗产等实物对象。表4列出了上述4个本体应用纲要的领域模型、核心类和所复用的词表。

在设计异构资源的本体应用纲要时，为了既保持术语词表定义的一致性同时又兼顾应用需求的特殊性，需遵循“术语复用一致性与应用纲要差异性原则”。对复用的外部术语词表保留其原始定义，对自定义的术语词表也在不同的应用纲要中保持一致。在特定的应用纲要中，同一个属性，其域和范围是一致的，但用于描述的类和预期取值及取值约束则根据具体的应用需求可能会有所不同。

由于本体应用纲要方法的灵活性和易扩展性，可以先后依次为不同的资源类型设计本体应用纲要，同时保证知识模型统一性；根据本体应用纲要的定义，采用关联数据和知识图谱技术对遗留系统中不同编码格式的元数据记录进行清洗、规范和丰富，统一转换为RDF序列化格式，则保证了知识表示的一致性。

4.2

从应用纲要到一体化本体设计

家谱、古籍、手稿档案、电影、音乐、老照片、老建筑、物质文化遗产等资源在各类文化记忆机构中具有典型性，在杨.阿斯曼的文化记忆理论中，被统称为文化记忆资源（CulturalHeritageMaterial），是以一定的形态固化下来、可长时间传承、大范围传播的文化记忆载体[23,24]。在上述各种特定文化记忆资源的本体应用纲要中，存在着共通的部分，主要是客观世界中存在过的各类实体和人们对客观实体的主观认识而形成的概念体系，其中实体包括人（Person）、机构（Organization）、地点（Place）、时间（Time）、事件（Event）、物体（PhysicalObject），概念体系则具体表现为各种领域的分类主题词表（Subject）。无论这些文化记忆资源以什么形态存在，都是客观的实体和主观的概念体系承载于一定载体上的文化记忆，单独的文化记忆资源反映了某个实体或概念的某一个侧面，要尽可能地完成某个实体或概念的完整拼图，需要尽可能全面地将各类相关的文化记忆资源联结在一起。例如要构建“胡适”人物实体的完整知识图谱，需要为与之相关的每一个文化记忆资源（家谱、书报刊作品、照片、音视频等）建立与“胡适”实体的关联，这样就能以“胡适”为中心，集中类型多样、位于不同文化记忆机构中的文化记忆资源。

本体应用纲要弥合了模型和格式不一致导致的信息孤岛之间的沟壑，而在本体应用纲要基础上建立一体化本体，对关于实体和概念体系的特征和关系以统一的知识模型和一致的知识表示形式化之后，构成了高层抽象的、共用共享的知识层，为多种类文化记忆资源在互联网和机器之间的知识融通奠定了基础。图3表示了跨越多种文化记忆资源类型的一体化本体知识融通模型，旨在厘清各类实体和概念体系与文化记忆资源之间的关系，这样的关系高度抽象，且与具体的资源类型无关。在为特定资源类型设计本体应用纲要时，有意识地遵照此模型设计出表示资源与实体和概念之间关系的属性，并在著录中实施，就能保证不同的资源尽可能地与相同的实体和概念建立关联。同时为同类实体和概念体系建立独立的知识图谱（知识库），作为各类文化记忆资源的链接中心。例如上海图书馆的人名规范库就是一个关于人物的独立实体库，其中的每一个人物实体都尽可能地与上海图书馆的家谱、古籍、手稿档案、电影、音乐、老建筑等文化记忆资源建立关联。

一体化本体的设计旨在为不同的应用纲要在类的复用和继承上保持逻辑上的一致性，图4展示了家谱、手稿档案、古籍、上海记忆等应用纲要在类的复用和继承关系。其中人（shl:Person）、姓氏（shl:FamilyName）、机构（shl:Organization）、地点（shl:Place）、时间（shl:Temporal）、事件（shl:Event）等实体类为所有应用纲要共用，每种实体本身也有自己的应用纲要，例如：描述人及与之相关的生平大事、任职经历、亲属关系和社会关系的人物本体应用纲要，描述机构及其沿革和机构间关系的机构本体应用纲要，关于地点及其历史变迁的地点本体应用纲要，描述各类历史纪年的时间本体应用纲要，描述事件及其涉及到的人物、时间、地点的事件本体应用纲要，以及揭示各种古籍分类体系的分类法本体应用纲要等。

4.3

一体化本体的扩展

灵活性和可扩展性是本体知识组织方法的重要优势。一体化本体的扩展包括两个层面，一是在统一的知识模型下为更多种类的文化记忆资源设计新的应用纲要，二是在原有的应用纲要上进行扩展。前者将更多的文化记忆资源纳入统一的知识模型中，后者则是对已有本体应用纲要的进一步完善。

上海图书馆人名规范库的人物本体应用纲要是在设计家谱、手稿档案、古籍、上海记忆本体应用纲要的过程中逐步完善而成。通过对人物本体应用纲要的扩展，还实现了人名规范库和中国历代人物传记资料库（CBDB）的跨领域知识融通。人名规范库作为图书馆界的规范控制方法与关联数据和知识图谱技术相结合的成果，最初的目的是为了实现互联网上的人名规范控制，但是在“数据基础设施”的建设中，逐步发展成为超越于资源类型的知识共享层和各类文化记忆资源的知识链接中心。人名规范库

中的人物主要来源于资源整理和编目过程中产生的人名规范档，选取范围是文化记忆资源的责任者和内容中涉及的重要人物，除了个人的籍贯、生卒年、字号排行等基本信息之外，还与大量现存于文化记忆机构中的各种文化记忆资源建立了关联。而CBDB则是由领域研究者从历朝历代的人物传记资料中整理出来的用于历史人文研究的结构化关系数据库，包含人物生平大事、任职经历、丰富的社会网络关系等。对于数字人文研究者来说，人名规范库和CBDB各有侧重，可以互相补充和丰富。为了将CBDB的数据与人名规范库融合，笔者对原有的人物本体应用纲要进行了扩展，通过自定义任职事件类（shl:OfficialEvent），使之继承原有的“事件（shl:Event）”类，用来描述CBDB的任职经历数据，通过自定义“关系类（shl:Relationship）”类，用来描述CBDB的社会网络关系数据。图5以苏轼为例，展示了CBDB的关系数据库格式的数据是如何融合到人名规范库的本体模型和关联数据之中。

5 结语

GLAM领域的文化记忆机构长期致力于文化记忆资源采集加工、知识组织与传播传承、跨机构资源共建与数据共享，一贯重视制订和遵循一定的领域标准规范进行数据加工，积累了大量高度结构化的元数据和规范数据记录。然而数字人文研究对跨机构、跨领域、跨网络、跨平台的数据融合和知识融通提出了更高的需求。由于GLAM领域的元数据主要描述的是文化记忆资源，而本体描述的是文化记忆资源中蕴含的知识。知识由实体和概念体系组成，是资源之上的抽象层，可作为资源之间链接的桥梁；知识是领域之上的共享层，可作为领域之间融通的中介。因此，基于本体的知识组织方法为数字人文系统提供了直接操控知识的工具，通过对知识的管理、检索、获取，来达到对不同领域不同资源的管理、检索和获取的目的。

然而文化记忆资源的多样性和数据格式的异构性，为实现跨领域跨机构的知识融通的同时还能满足特定资源的个性化描述与揭示带来了困难。笔者在“元数据应用纲要”的基础上提出了“本体应用纲要”的概念及其设计原则和流程，以规范特定资源的本体应用纲要的设计同时满足其个性化需求；之后在家谱、手稿档案、古籍、上海记忆本体应用纲要上抽象出文化记忆资源的“一体化本体”知识融通模型，以保证不同资源类型的本体应用纲要在知识建模上的统一性，从而促进跨机构的知识融通；进而通过类和属性的复用和继承，以方便地对一体化本体进行扩展，甚至可与更多人文领域的知识进行融合，从而促进跨领域的知识融通。总之，元数据应用纲要为文化记忆资源的知识融通提供了结构化数据储备，本体应用纲要兼顾了特定资源类型的个性化需求，一体化本体的知识融通模型则可保证知识建模的统一性和知识表示的一致性。“本体应用纲要”设计方法和“一体化本体”知识融通模型与关联数据和知识图谱技术结合后，为跨机构跨领域的知识融通提供了方法和路径。

鉴于目前“应用纲要”的方法和技术已在国外图书馆和语义网领域有着深入的应用，本文提出的“本体应用纲要”设计方法和“一体化本体”融通模型，已经应用于上海图书馆的古籍、档案、民间文献和实物等资源的描述与揭示，在与其他GLAM机构的古籍家谱等资源的整合中，构建了人名规范库、地理名词表、历史纪年词表、历史文化事件知识库、优秀历史建筑知识库等与特定资源类型和数据格式无关且广泛关联的实体层（人、地、时、事、物）和概念体系层（主题、分类），已成为链接不同机构不同资源的数据链接中心和融通不同领域知识的知识共享中心，在跨机构和跨领域的知识融通上起到了一定的作用。建议档案馆、博物馆等文化记忆机构尝试使用“本体应用纲要”的方法和关联数据、知识图谱技术，融入并完善“一体化本体”知识融通模型，以实现跨机构的开放互联和跨领域的知识融通。

支撑数据

数据开放获取地址:http://data.library.sh.cn。

参考文献

*本文原载于《图书情报知识》2021年第1期53-65页。