基于书目框架(BIBFRAME)的家谱本体设计*

夏翠娟，刘　炜，张　磊，朱雯晶

摘　要　文章针对图书馆的家谱系统大多只提供基于关键词匹配的字段检索，主要面向文献特征而不能深入揭示内容的问题，在对家谱信息系统及相关技术现状进行文献调研的基础上，论述了上海图书馆基于书目框架模型来设计家谱本体的起因、过程、方法和成果；利用本体建模方法设计了上海图书馆家谱本体及其应用、扩展和重用；得出了基于书目框架设计的家谱本体既能揭示家谱资源的文献特征和内容属性，又能增强内容之间语义关联的结论。

关键词　家谱本体　书目框架语义网　关联数据

引用本文格式　夏翠娟，刘炜，张磊，等. 基于书目框架（BIBFRAME）的家谱本体设计[J]. 图书馆论坛，2014（11）：5-19.

A Genealogical Ontology in the Form of BIBFRAME Model

XIA Cui-juan，LIU Wei，ZHANG Lei，ZHU Wen-jing

Abstract　Most of the existing genealogy systems in libraries only provide keyword search based on fields and subfields. They are not able to reveal the content relationships among the data entities such as Family，Person，Place，Event etc. effectively in the genealogical documents. After making a literature review，this paper studies on how to resolve these problems by using semantic web and linked data technologies. It mainly discusses the design of genealogical ontology based on bibliographic framework，which is a data model for bibliographic description and has been developed to replace MARC using Linked Data technology. The implementation，extension and re-use of the genealogical ontology are also discussed.

Keywords　genealogical ontology；bibliographic framework；semantic web；linked data

0　引言

书目框架(BIBFRAME)是美国国会图书馆牵头开发的下一代书目数据格式标准，也是该开发项目的简称。自 2011 年 5 月起，美国国会图书馆联合大英图书馆、德国国家图书馆等 6 个图书馆，请 DC 元数据的发明人之一，也是语义万维网技术的倡导者 Eric Miller 领衔，正式启动“书目框架计划”。该计划的主要目标是设计一套互联网时代的书目数据标准，用以取代 MARC，并能为图书馆、档案馆、博物馆、美术馆等“人类文化记忆机构”共同使用[1]。经过 3 年多的开发，书目框架模型基本成型，各相关规范的文本编写接近尾声。目前其官方网

*本文系国家社科基金青年项目“W3C 的 RDB2RDF 标准规范在关联数据服务构建中的应用”（项目编号：13CTQ008）和国家社科基金一般项目“关联数据的理论与应用研究”（项目编号：11BTQ04）研究成果之一

站(http://www.loc.gov/bibframe/)发布的成果包括书目框架模型(BIBFRAME Model)、术语词表(BIBFRAME vocabulary，包含 300 多个术语，并还在根据需要增加和修订)、BIBFRAME 纲要(BIBFRAME Profile，对于各类“社区”应用书目框架的进一步限定或扩展的规定)、书目框架权威档(BIBFRAME Authorities)、关系描述(BIBFRAME Relationships)以及 MARC 数据转换为 BIBFRAME 格式的工具、书目框架编辑器(BIBFRAME Editor)的演示平台等，内容非常丰富；但行百里者半九十，尚有一些关键细节还没有定论，如对书目框架的形式化表达和书目数据的 RDF 序列化规则等方面还有大量的工作要做，特别是对如何保留或有没有必要保留那些基于 AACR2 或 RDA 的编目规则而得到的大量丰富而微妙的语义，正在进行激烈的论辩。

家谱是一类记载具有血缘关系的家族世系繁衍情况和重要人物及事迹的历史文献，是研究人文历史和地域文化的重要资源。上海图书馆是全世界收藏中文家谱(原件)数量最多的机构。为了更好地保护和庋藏这些资料，上海图书馆在过去 10 多年一直在进行家谱的整理和数字化工作，初步建立了包含 1.8 万余种家谱的影像资源库，以图书馆人熟悉的 MARC 格式作为数据检索和交换格式提供服务。近年随着“数字人文”研究的兴起和各类相关工具平台的建立，基于文献的揭示方式难以满足学者进行深入研究的需要。比如，家谱中包含丰富的人、地、时、事、机构及相互关系等，都不是基于 MARC 的系统所能描述和揭示的，还必须进一步进行基于内容的深度加工和揭示，并提供灵活的、多维度的展示和操控工具，才能使数字家谱得到更好的利用。

语义万维网技术尤其是关联数据技术为上述需求提供了可行的方案。书目框架就是该技术在图情领域的最新应用，正好能为重组家谱资源、重构家谱服务系统提供新的解决方案。书目框架是基于关联数据技术框架设计的。关联数据是语义万维网的轻量级实现方式，它植根于现有的 Web 基础技术：用 HTTP URI 来标识数据，使 URI 不仅作为事物的名称，同时兼作存取地址；以服务器对不同请求的响应来区分信息资源或非信息资源；采用 RDF 模型作为描述世间万物及其相互关系的基本结构，在此基础上可以利用万维网本体语言(OWL)建立更为复杂的领域知识模型，为更广泛的基于机器理解的语义互操作奠定了基础[2-3]。

知识本体给数据赋予了语义，关联数据技术以标准的格式为数据编码使得机器能够理解语义并处理数据间的关系。本文提出采用语义万维网技术来建设新的家谱知识库系统，设计一个向下兼容、易于扩展、便于重用和共享、支持家谱数据重组和知识建模的家谱知识本体，这是首要的工作。设计知识本体的一个重要原则是尽量复用已有的本体模型和术语词表。本文在文献调研、家谱领域现有案例分析以及技术现状研究的基础上，基于书目框架模型，复用书目框架术语词表中的术语，设计了上海图书馆家谱本体，并采用书目框架应用纲要来规范家谱本体的应用和实施，这是利用语义万维网技术改造图书馆传统资源的组织方式，以提升服务效果的一种尝试，也是对正在发展之中的书目框架应用于中文环境的试验和检测。

1　家谱信息系统建设及研究现状

1.1　现有的家谱应用系统和技术

家谱收藏机构主要是图书馆和教会、宗亲会等机构。在国外，家谱收藏机构有美国犹他家谱研究学会、日本国立国会图书馆等；在我国港台地区，台湾“故宫博物院”和台北“国家图书馆”收藏家谱较多，香港大学图书馆也有少量收藏；我国大陆家谱收藏和研究机构主要有中国国家图书馆，上海图书馆等几个大型的省级公共图书馆，以及少数高校图书馆。目前主要的家谱应用系统有犹他家谱研究学会的家谱检索中心(FamilySearch.org)、日本国立国会图书馆的东洋文库、中国国家图书馆的“中

华寻根网”、上海图书馆的家谱数据库、台湾地区家谱联合目录数据库、《四库全书》等大型数字化古籍数据库中的家谱资源库等。王昭[4]和毛建军[5]对上述家谱收藏机构和家谱应用系统进行了介绍分析。

日本国立国会图书馆的东洋文库、中国国家图书馆的“中华寻根网”、上海图书馆的家谱数据库均采用题名、著者、姓氏、居地、名人等字段进行检索，是基于字段关键词匹配、面向家谱文献资源的检索系统。犹他家谱研究学会的 FamilySearch.org 不仅可以根据文献的收藏地、类型、批次号码和缩微胶卷编号来查询家谱资料，还可根据姓氏和名字、生平事迹(出生、结婚、居所、死亡等)、配偶或父母关系来查询。国外还有 Ancestry.com 和 WeRelate 等家谱网站，与 FamilySearch 一样，允许用户自行创建家族树，上传家族照片和撰写人物生平大事，甚至多个不同用户可共同维护一棵家族树。

国外家谱领域应用较为广泛的技术标准是 GEDCOM，较有影响的家谱概念模型是 GENTECH。GEDCOM 是用于在不同的家谱软件之间交换数据的家谱数据交换标准，最开始是为耶稣基督后期圣徒教会(The Church of Jesus Christ of Latter-day Saints)的需求设计，也被美国犹他家谱研究学会采用。它不是一个数据模型，可看做是用于家谱数据的文本标记语言。GEDCOM 文件是包含家谱文献元数据记录的纯文本，其结构适合于 20 世纪 90 年代的技术环境。Campanya Artes Joan[6]指出：在目前的环境下，它有以下几个弊端：专用的格式不利于进一步发展；标准的定义不够严谨，在应用过程中容易产生分歧；数据冗余导致不一致性；没有足够的灵活性来适应不同的文化环境，如人名、地名的定义和描述，只能用于家谱领域，无法与其他领域进行数据交换。GENTECH 是一个家谱概念模型，源于一个研究者之间的合作项目，只在 2000-2004 年间延续了很短的时间，但得到美国全国宗谱协会(U.S. National Genealogical Society)的关注。虽然它没有具体的应用实施方案指南，但常被作为许多相关应用的参考。GENTECH 在某种程度上提供了一种处理复杂问题的解决方案[7]，比如不同历史时期同一地理位置具有不同的地理名称的问题；另一方面，该模型将所有与人有关的信息关联起来，比如机构、历史事件、家族活动，还提供将初始数据表达成为具体应用所需的不同形式(文档、记录、文件)的灵活性和可扩展性。由于GENTECH 没有成为被广泛接受的标准规范，没有得到应用和推广。GEDCOM 和 GENTECH 主要是为欧美家谱而设计，在我国少见应用。

21 世纪初，W3C 推出诸如 XML 超文本标记语言，GEDCOM 为适应这个趋势进行升级，GEDCOM6.0 版也叫 GEDCOM XML。其它基于 XML 格式的家谱标记语言 GedML、EeniML、GenXML，与 GEDCOM 一样，只有少数机构在使用。随着语义万维网概念的提出，W3C 又推出资源描述框架(RDF)、知识本体语言(OWL)等语义万维网相关标准规范。RDF/XML 作为 W3C 的推荐标准和语义万维网技术的基础，可被大部分机器语言识别和处理，已被广泛应用于多种不同的领域，有利于跨领域的共享和重用。Jay Askren 开发了传统的 GEDCOM 格式转换为 RDF/XML 格式的工具，以证明 RDF 的广泛适应性[8]。语义技术作为历史研究的工具得到重视，Albert Meroño-Peñuela 对基于语义技术的历史研究方法作了调研，其中涉及家谱研究[9]。关联数据作为语义万维网的轻量级实现方式也受到关注，Josh Hansen[10]论述了利用关联数据技术来实现家谱数据全球共建共享的可行性和方法，其中提到了基于关联数据技术的一个家谱数据集 John Goodwin’s Family Tree[11]，该数据集已在最大的关联数据集注册中心 thedatahub.org 注册。

1.2 现有的家谱元数据方案和家谱本体

在资源组织上，图书馆习惯于将家谱作为一种历史文献资源来保存和处理，主要集中在对家谱文献的整理和元数据著录上，过去大多

利用图书馆编目系统著录，采用 MARC 数据格式。近年开始采用 DC 元数据标准来为家谱资源设计元数据方案，尤其是在我国，如科技部科技基础性工作专项资金重大项目——我国数字图书馆标准与规范建设的家谱元数据规范子项目的成果：张秋芳等人的《家谱描述元数据规范》[12]；国家数字图书馆工程标准规范项目的成果：赵亮等人的《国家图书馆家谱元数据规范与著录规则》[13]，上海图书馆参与了这两个项目。在这两个项目的元数据方案中，元数据元素大都包括题名、卷数、修撰者(著者)、版本、谱籍地、堂号、始祖、始迁祖、收藏地、提要等信息。2000 年由上海图书馆牵头，犹他家谱研究学会以及我国台湾、香港的家谱收藏机构参与整理的《中国家谱总目》[14]是迄今为止收录我国家谱最多、著录内容最为丰富的一部专题性联合目录，基本采用上述元数据元素。

近年逐渐出现基于知识本体的解决方案，国外有多篇文献记载家谱本体的设计。2005 年荷兰一家为图书馆、档案馆、博物馆提供咨询服务的公司 Ivo Zandhuis[15]论述了家谱本体的设计，定义了一套术语词表，以 RDF/XML 格式在 Web 上发布。美国杨百翰大学(BYU)的 Charla Woodbury 和 David W.Embley 在探索中记载了设计的家谱本体和基于本体进行逻辑推理和知识挖掘，处理同一人多名的方法[16]。Josh Hansen 阐述了基于关联数据技术的家谱本体设计思路，发布了一个家谱本体术语词表(http://purl.org/gen/0.1#)。在我国，上海交通大学的陈艳以上海图书馆的家谱为例论述了中国家谱本体的构建方法、过程和结果[17]，武汉大学的董慧等在 2008 年 IEEE 大会上介绍了基于本体的家谱知识建模方法[18]，遗憾的是没有公开发布家谱本体术语词表。

与家谱信息中包含多种实体相关的知识本体有用于人和机构的本体 FOAF[19]、关于人与人之间关系的本体 Relationship[20]、地理本体 GeoNames Ontology[21]、时间本体 TimeOntology[22]、事件本体 Event[23]，Albert Meroño-Peñuela 对上述本体作了调研[9]。这些本体为解决家谱中的具体问题提供了建模方法上的参考，且其词表以 RDF/XML 格式发布，其中的术语可被方便地重用在家谱的本地应用系统。在图书馆中，家谱作为文献的特征仍然需要得到充分揭示，相关的本体有欧洲数字图书馆数据模型(EDM)[24]、OCLC 的 Schema.org 书目扩展 SchemaBibEX[25]，以及美国国会图书馆的书目框架，与前二者相比，书目框架明确以替代 MARC 为目的，不仅仅是一种书目格式，而是一个从模型到词表、到实现技术的系统性框架。书目框架能兼容 RDA、FRBR 等已有的标准，也支持与 SchemaBibEX 甚至档案界 VRA 模型的互操作，既能够深度描述资源的文献特征，也能描述人、地、时、事等内容特征，虽然尚有诸多细节有待讨论，但仍被寄予厚望。

2 上海图书馆家谱本体模型和术语词表的设计

目前的家谱信息系统大致可以分为两类：一是以家谱文献为主要管理对象，二是以家族世袭人物关系及其相关事迹记载为主要对象。当然，这两类信息系统经常无法截然分开，前者必然会涉及家谱对内容的描述，比如始祖、始迁祖、宗族名人、迁徙地；而后者也离不开家谱文献，也是通过对家谱记载或修谱时描述而进行记录。

上海图书馆已有的家谱系统以家谱文献为管理对象，采用对文献进行著录的一整套元数据元素集，以 MARC 为数据格式，可通过题名、姓氏、居地、堂号、著者、名人、丛书、索取号等与家谱文献相关的字段进行检索，在家谱阅览室可以查看扫描的影像文件。这种仅仅以文献方式建立的信息系统在很多时候无法满足用户的查检需求，最大的问题是缺乏规范控制，对于姓氏、年代、人名、地名等所有字段都只能采用关键词(自由词)匹配而不是概念匹配，缺乏必要的准确性，极大地影响了查全率和查准率，而且缺乏聚类功能、关联关系的发

现等。这些缺陷正好都是目前关联数据技术的强项，也是上海图书馆要以尚未开发完成的书目框架模型来建立家谱信息本体的主因，希望能够兼顾家谱文献管理和内容揭示两方面需求，使图书馆的信息系统由于应用了语义技术，而能够为更多的人所利用。

2.1 书目框架的核心模型和本体词表

本文中所说的知识本体(有时简称本体)，是专指对领域知识进行抽象，建立一定的概念模型，并使计算机能够“理解”这个模型的一种形式化知识表达工具。知识本体常常表现为一套体系化的术语词表及其相互之间关系描述，并以一定的机器语言进行编码而得到的代码体系。比如，传统分类编目工作常用的分类法和主题词表等，如果以 SKOS 这种专门的、基于 RDFS 的编码规范进行编码之后，所形成的知识体系就可看成是一种本体。知识本体应包括每一个术语的明确定义及其关系(比如叙词表种的用、代、属、分、参之类的关系)，术语分为类(Class)和属性(Property)两种，类是对同一类实体对象的抽象，属性是对类的各种特征的抽象，用于表示类与类之间的关系。

书目框架是图书馆领域一个最新的本体模型，它由许多不同的实体类和属性构成，类和属性的定义及取值都在书目框架术语词表(BIBFRAME vocabulary)中规定。书目框架模型[26](见图 1)包含四大类：创造性作品(Work)、实例(Instance)、规范(Authority)、注释(Annotation)，其中与文献相关的是作品和实例，与内容相关的属性属于作品，与格式和载体相关的属性属于实例。这与书目记录的功能需求(FRBR)模型的四大类相比更为简洁，作品对应 FRBR 模型中的作品(Work)和内容表达(Expression)，实例对应着 FRBR 模型的载体表现(Manifestation)，而 FRBR 中与馆藏复本相关的单件(Item)则作为书目框架的注释(Annotation)的一个子类。注释体现了书目框架模型的开放性，在注释模型中，容纳馆藏相关的本地信息，可以将各种互联网资源如书评、评分等信息与书目数据相关联。FRBR 第二组实体人、机构等规范控制相关的数据在书目框架中属于规范(Authority)，规范提供一个轻量级的规范控制层，可利用已有的规范词表如 VIAF、LCSH 等，使 Web 级的规范控制更为有效。基于书目框架模型设计家谱本体，可以将家谱数据中的内容和载体明显地区分开，并利用书目框架的规范控制方法，

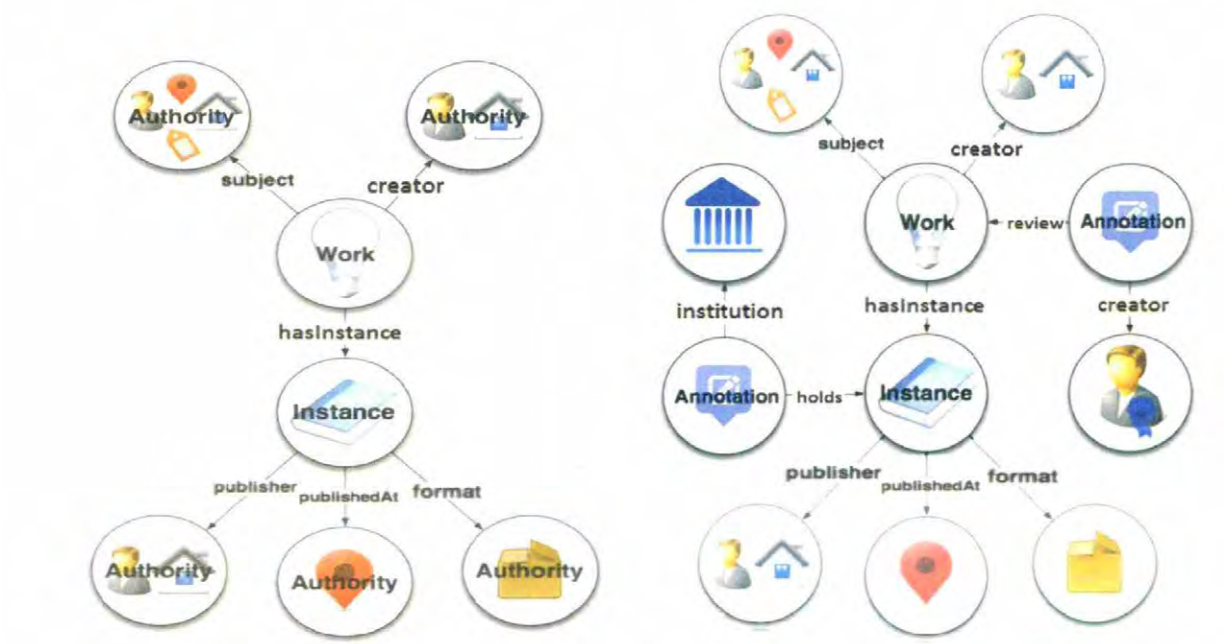

图 1 书目框架的核心模型和注释模型

实现基于 Web 的规范控制，利用注释模型引入更多的开放资源，补充家谱知识库的不足。

到目前为止，书目框架术语词表共定义了 338 个术语，除了明确定义核心模型的四大类外，一些与四大类相关的其他资源也被抽象为与这四类同级的资源类，都作为 bf:Resource 类的子类，比如事件(Event)、关系(Related)、题名项(Title)、标识符(Identifier)、语种项(Language)等，根据关联数据的原则，这些在 MARC 记录中以文本出现的字段值在书目框架中作为资源对象来处理。家谱中的各类数据实体包括文献相关的类，如题名、责任者、载体项、出版项等，以及可用于家族信息建模的类，如人、家族、机构、地、时、事等均可在书目框架中找到对应的类——bf：Person，Organization，bf：Place，bf：Temporal，bf：Family，有丰富的属性来表达类与类之间的关系。这样原 MARC 记录中作为文本串的数据可以作为资源对象，利用明确定义的属性来表达对象之间的关联关系，为数据赋予语义，便于机器处理和跨系统的互操作。该术语词表已用 RDF Schema 编码，提供 RDF/XML 格式的文件下载。

上海图书馆大量已有家谱数据是 MARC 格式，而书目框架的目标在于取代 MARC，并非抛弃 MARC。大量的 MARC 格式的数据是图书馆的宝贵财产。新的书目格式必须兼容旧格式，使已有数据能够顺畅地转换为新格式。BIBFRAME 的核心数据模型和本体词表全面考虑了 MARC 格式的兼容性，且项目组正在开发 MARC 转换为 BIBFRAME 的工具和平台。基于书目框架来设计家谱本体，在系统实现时可以利用这些工具、平台，借鉴其方法。

2.2 从家谱元数据到家谱知识本体

上海图书馆现有的家谱数据库已有一套元数据方案，这决定了数据库中的元数据记录的结构。知识本体的设计必须考虑容纳现有的数据项，基于现有数据结构来厘清数据之间的关系。知识本体是元数据方案的立体化[27]，有哪些元数据元素决定着需要设计哪些类和属性。元数据方案是平面的，而知识本体则是厘清了元素所描述的类(Class)，定义了类与类之间的关系，以属性(Property)来明确表达这些关系而形成的立体网状模型[28]。在设计本体时，一个原则是尽可能地复用已有本体的类和属性，如果已有本体中的类和属性不足以表达具体应用领域中的数据实体及其关系，就需要自定义新的类和属性。上海图书馆的家谱本体需建立在上海图书馆的家谱元数据方案之上，表 1 是上海图书馆家谱元数据元素与书目框架术语词表中类和属性的对应。从家谱元数据中可以发现家谱资源与图书馆其他资源相比的共性和特殊性。共性表现在题名项、责任者项、出版项、载体形态项、馆藏项等文献特征，这在书目框架术语词表中有足够的类和属性与之相对应。特殊性表现在和家族相关的属性如始祖、始迁祖、散居地等，人的属性如姓、名、字、号、兄弟排行等属性是家谱甚至是我国家谱所独有的信息，书目框架的类和属性不足以描述这些特有属性，现有的家谱本体以及应用最为广泛的描述人的本体 FOAF 也没有相应的属性来描述这些特性，因而需要自定义家谱资源专有的类和属性。在自定义类和属性时，尽量用继承的方式继承书目框架已有的类及其属性，这样就能继承父类中已有的属性，并保证与书目框架兼容。

家谱中的迁徙信息一般由人(始祖或始迁祖)、地(原居地和迁居地)、时(何时迁往何地)三要素构成，因此被作为事件(bf：Event)来处理。始祖、始迁祖、支祖、房祖、名人等人有所处时代、原居地、迁居地、名、字、号、排行等特性，可以用一个特定的类及其属性来建模。始祖、始迁祖、支祖、房祖、名人、散居地等属于某个家族的信息，可用“家族”类来建模。因而自定义了三个类： shlgen：Family(家族)；shlgen：Person(人)；shlgen：FamilyName(姓氏)。“shlgen”是上海图书馆家谱本体命名空间的前缀，带有该前缀的类和属性即为自定义的类和属性。其中 shlgen：Family 继承 bf:

表 1　家谱元数据与家谱知识本体的对应关系

家谱元数据		家谱知识本体	
元数据项	元数据元素	类	属性
标识	标识符	bf：Identifier (标识符)	bf：identifierType(标识符类型) bf：identifierValue(标识符值)
题名项	书名	bf：Title(题名)	bf：titleType(题名类型) bf：titleValue(题名值) bf：workTitle(作品题名)
	书名来源		
	异书名		
	异书名来源		
地理信息	谱籍	bf：Place(地点)	bf：place(地点)
时间信息	时间范围	bf：Temporal(时间)	bf：temporalCoverageNote(时间)
责任者项	责任者	shlgen：Person(人)	shlgen：familyName(姓) shlgen：ming(名) shlgen：temporal(时代) bf：role(角色)
	责任者时代		
	责任方式		
	其他责任者		
	其他责任者时代		
	其他责任者责任方式		
版本项	版本		bf：editon(版本)
出版项	年代	bf：Provider(提供者)	bf：providerName(提供者名称) bf：providerDate(提供时间) bf：providerPlace(提供地点)
	堂号		
	镌刻者		
载体形态项	装订	bf：Category(类别)	bf：extent(容量) bf：category(分类) bf：carrierCategory(载体类型) bf：categoryType(类别) bf：categoryValue(分类值)
	数量		
附注项	谱载内容	bf：Summary(提要)	bf：hasAnnotation(注释)
馆藏项	收藏者	bf：HeldItem(馆藏单件) bf：Organization(机构)	bf：heldBy(收藏者) bf：shelfMarkDdc(杜威十进制分类号)
	索书号		
家族信息	迁徙信息	bf：Event(事件) shlgen：Family(家族) shlgen：Person(人) shlgen：FamilyName(姓氏)	bf：eventAgent(事件主体) bf：eventDate(事件发生日期) bf：eventPlace(事件发生地点) shlgen：migration(迁徙) shlgen：ancestor(始祖) shlgen：branchAncestor(支祖) shlgen：familyAncestor(房祖) shlgen：notableAncestor(名人) shlgen：firstMigratedAncestor(始迁祖) foaf：name(名称) shlgen：family(家族) shlgen：temporal(时代) shlgen：originalLocality(原居地) shlgen：locality(迁居地) shlgen：otherLocalities(散居地) shlgen：familyName(姓) shlgen：givenName(名) shlgen：courtesyName(字) shlgen：pseudonym(号) shlgen：orderOfSeniority(排行) shlgen：event(生平事迹)
	始祖		
	始祖时代		
	始祖原居地		
	始祖迁居地		
	始迁祖		
	始迁祖时代		
	始迁祖原居地		
	始迁祖迁居地		
	支祖 / 房祖		
	支祖 / 房祖时代		
	散居地		
	姓氏		
	名		
	字		
	号		
	行		
	名人		
	名人时代		

Family，shlgen：Person 继承 bf：Person，bf 是书目框架命名空间的前缀，带有该前缀的类和属性即为书目框架所定义。之所以要把姓氏 shlgen：FamilyName 也定义为一个类，是因为在家谱数据中，姓氏是非常重要的资源，上海图书馆的家谱数据中包括 335 个姓氏，张、陈、王、李、刘、吴等姓的家谱文献均在 500 种以上，而周、朱、徐、黄、杨、胡等姓也达数百种之多，冷僻姓氏有 90 余种。将姓氏作为资源对象来处理，有利于将关于姓氏的信息，如发源地、地域分布等数据结构化、语义化。

书目框架本体对家谱文献特征描述的类和属性较为充足，无需作进一步扩展。遵照书目框架的核心数据模型，将家谱分为作品、实例两个主要部分。家谱元数据中的题名项、责任者项、附注项以及其他与家谱文献内容有关的人、地、时、事、家族信息等属性归于作品，将与文献载体有关的出版项、载体形态项、版本项归于实例，而规范与注释都通过作品和实例各自的属性所定义的关联关系与作品和实例相关联，见图 2。

图 2　基于书目框架的家谱本体模型

2.3　家谱实体–关系模型

作品、实例、家族、人之间的实体关系可用实体关系图来分析。图 3 是与作品相关的类和属性关系图，图中圆角矩形表示类，用带箭头的有向线条表示属性，用直角矩形表示文本串(Literal)，子类用 rdfs：subClassOf 表示，类及其属性的域(Domain)和范围(Range)可从图中的有向线条及其起止点看出。比如，代表属性“bf：creator(责任者)”的有向线条从 bf：Work 类指向 bf:Agent 类，那么属性“bf：creator(责任者)”的域是类“bf：Work(作品)”，表示该属性是用于描述该类的，其范围是 bf：Agent 类，表示该属性的取值属于 shlgen：Person 类，而属性 bf：role(责任方式)的取值是一个文本串(Literal)，用直角矩形表示。

自定义的 shlgen：Family 类和 shlgen：Person 类是这样与 bf：Work 类发生关联的：作品的主题属性(bf：subject)的范围是 bf：Authority，而 shlgen：Family 类继承了 bf：Authority 的子类 bf：Agent 的子类 bf：Family(见图 5)，所以也继承了 bf：Authority，故可以将 shlgen：Family 作为作品主题的一种加以揭示。作品的责任者属性(bf：creator)的范围是 bf：Agent，而 shlgen：Person 继承了 bf：Agent 的子类 bf：Person(见图 6)，因而可以将 shlgen：Person 作为责任者的一种。

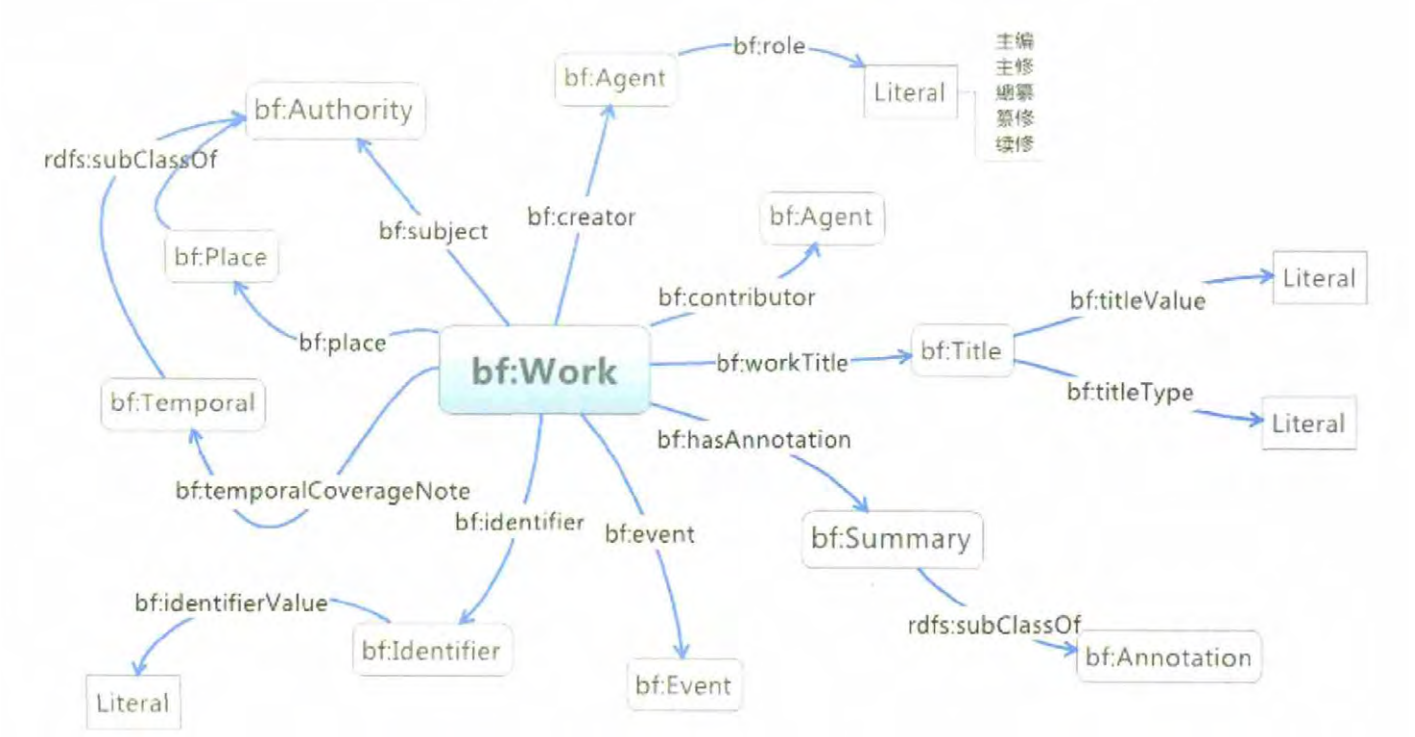

图 3　作品相关的类、属性及其关系

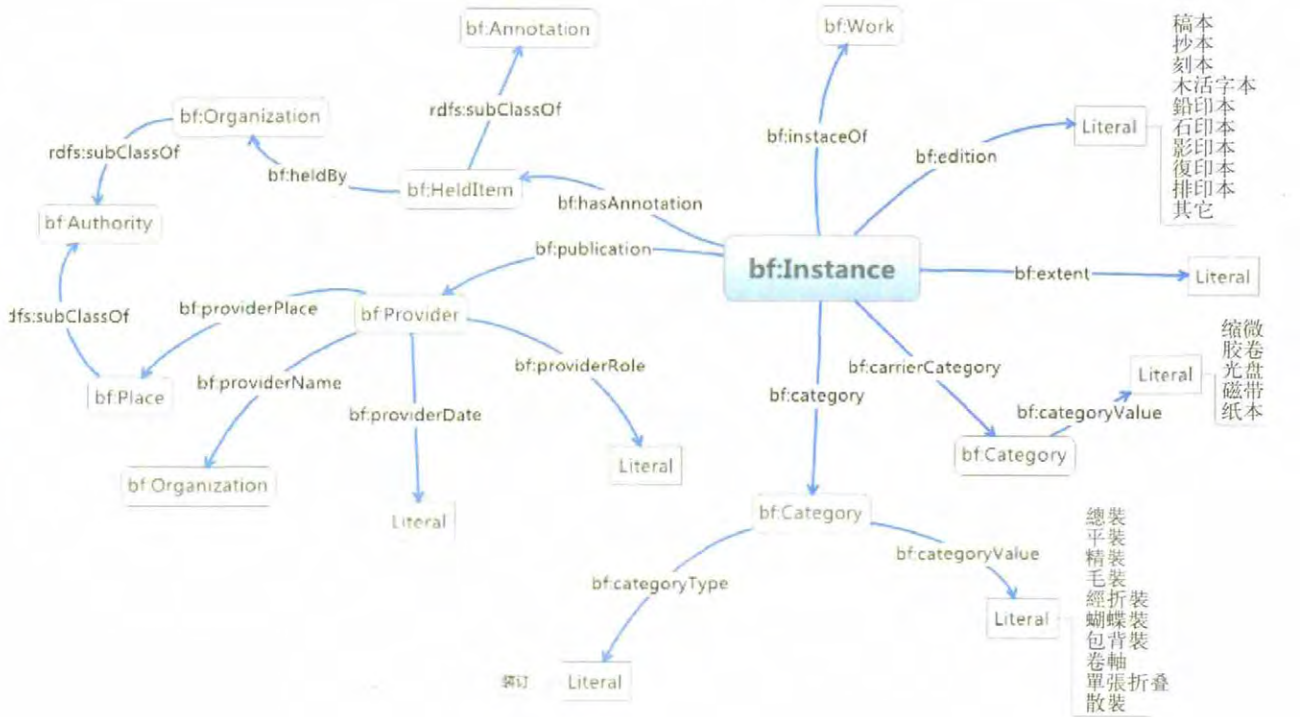

图 4　实例相关的类、属性及其关系

地点和时间通过属性 bf：place 和属性 bf：temporalCoverageNote 来与 bf：Work 发生关联，这两个属性的范围分别是地点(bf：Place)和时间(bf：Temporal)，都是规范(bf：Authority)的子类。图 4 中的收藏者属性(bf：heldBy)所指向的机构(bf：Organization)和出版地属性(bf：providerPlace)所指向的地点(bf：Place)也是如此。对注释(bf：Annotation)来说，作品的附注

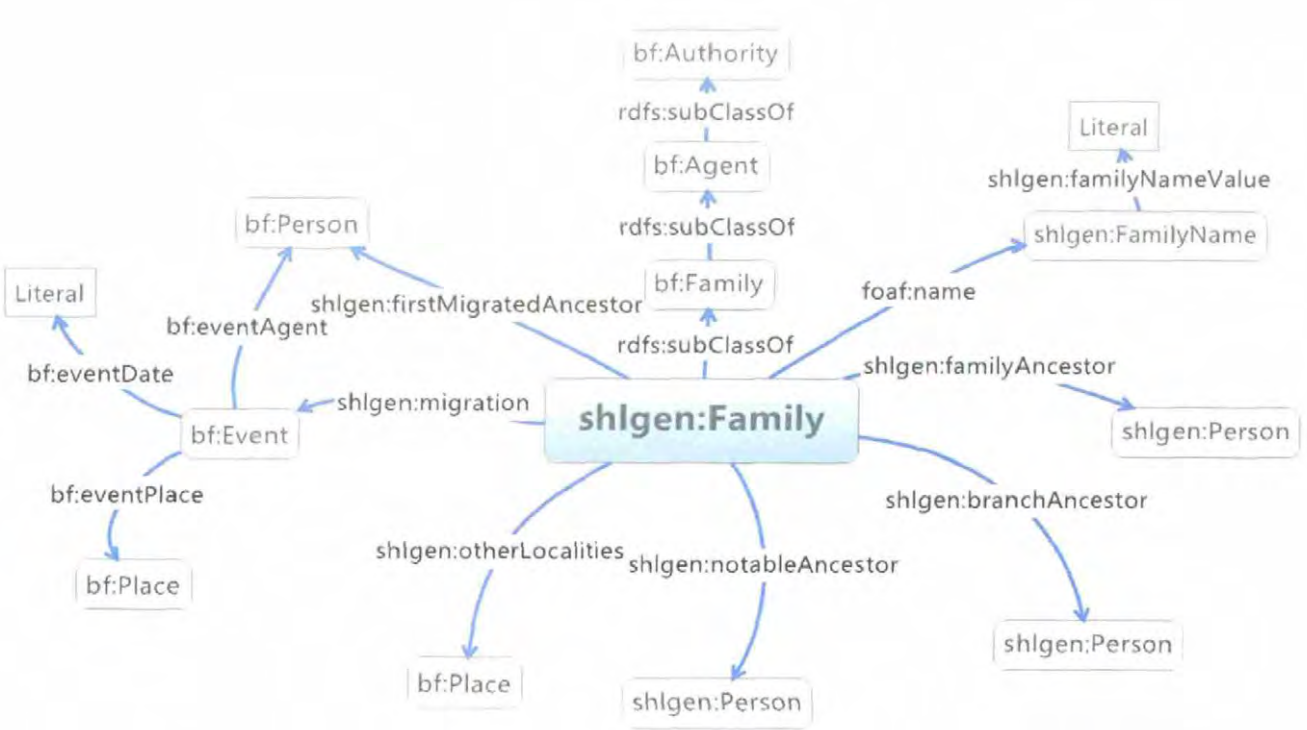

图 5　上海图书馆家谱本体中家族相关的类、属性及其关系

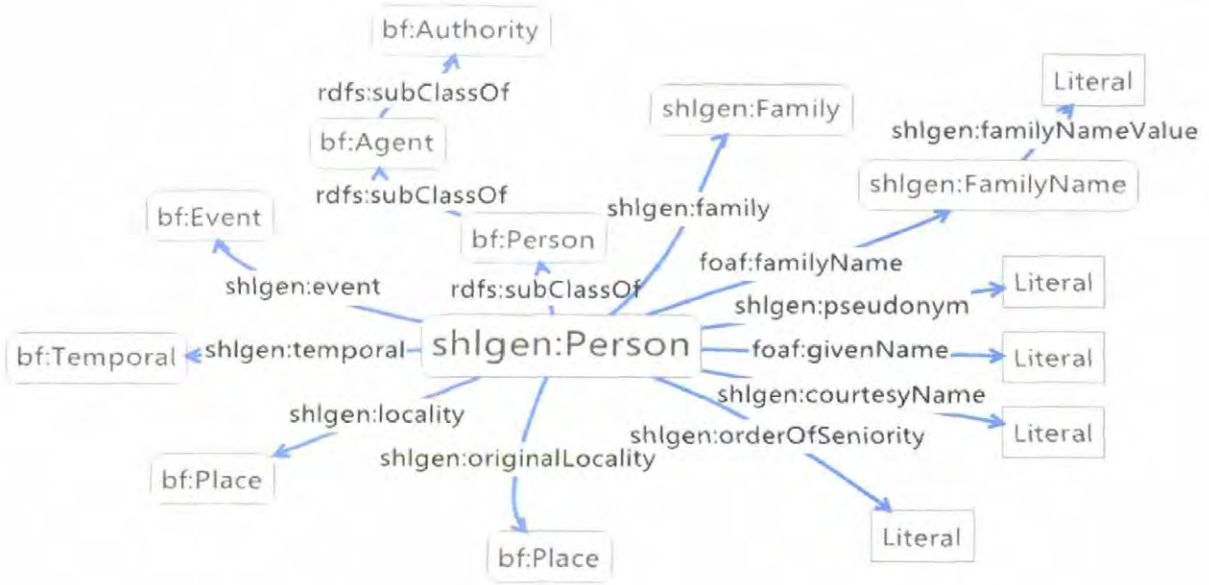

图 6　上海图书馆家谱本体中人相关的类、属性及其关系

(bf：Summary)(见图 3)是它的子类，实例的馆藏信息(bf：HeldItem)(见图 4)是它的子类 bf：HeldMaterial 的子类。

值得注意的是，对责任者和相应的责任者角色的对应处理，在书目框架里有两种方式：一是 bf：creator 直接指向责任者实体对象；二是 bf：creator 的范围是一个抽象的中间类 bf：Related(关系)，由“关系”类的属性 bf：relatedTo 来指向责任者实体对象，由 bf：relatedType 来表示相应的责任者角色，这里采用第一种方法，最新的 BIBFRAME 本体词表中也将 bf：creator 的范围定义为 bf：Agent 类。bf：Agent 是 bf：Authority 的子类，子类可以继承父类的属性，因而用从 bf：Authority 类继承过

来的属性 bf：role 来表示责任者的角色，其范围是一个文本串，取值约束定义为一个列表：主编、主修、总纂、纂修、续修(见图 3)。对取值约束的定义在“实例”的版本(bf：edition)属性和载体形态属性(bf：categoryValue)上也有体现(见图 4)。

在书目框架中，很多在元数据记录中取值范围为字符串的属性被作为实体对象来处理，如标识符、题名、版本项、载体项、出版者项。以题名为例，作品的题名属性 bf：workTitle 的范围不再是一个文本串，而是 bf：Title 类，该类的两个属性 bf：titleType 和 bf：Value 分别定义题名的类型(缩写、封面、书脊……)和值。对于上海图书馆家谱数据来说，当一个作品有多个书名时，用这种面向对象的方式更易于处理书名类型和值的对应关系。用 RDF 三元组表示如下：

作品 0010012——bf：workTitle——题名1
题名 1——bf：titleType——“卷端”
题名 1——bf：titleValue——“维扬安阜洲丁氏重修族谱六卷”
作品 0010012——bf：workTitle——题名2
题名 2——bf：titleType——“版心”
题名 2——bf：titleValue——“丁氏族谱”

家族 shlgen：Family、人 shlgen：Person、姓氏 shlgen：FamilyName 这三个类及其属性见图 5 和图 6 所示。

3 上海图书馆家谱本体的实施、扩展与重用

3.1 基于书目框架应用纲要的家谱本体实施

书目框架是一个试图兼容 MARC、RDA、VRA 以及未来可能出现的标准规范的框架，被设计成具有一定的灵活性和可扩展性，因而不能对具体领域的具体应用作出具体的规定。书目框架应用纲要是根据具体需求为领域本体的实施和应用在语法、用法甚至数据格式上作出明确定义的规范文档，它独立于书目框架模型和术语词表，由特定的应用领域自行维护，以适应具体的应用需求。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，可被机器处理，是抽象的本体到具体的应用系统之间的桥梁。书目框架应用纲要(BibFrame Profile)规范[29]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。应用纲要由“纲要定义(Profile Definition)”和多个“资源模板(Resources Templates)”组成，“纲要定义”声明了该应用纲要用于哪种，比如“专著”、“信函”等，“资源模板”规定具体应用纲要包含哪些类(如作品、实例、规范、注释)。一个“资源模板”包含多个“属性模板(Properties Template)”。属性模板定义一个类有哪些属性，各个属性的域和范围，以及属性的数据类型约束和取值约束。“纲要定义”“资源模板”和“属性模板”都有各自的元素来明确定义，比如“纲要定义”需由 identifier(应用纲要的标识符，机读)、Title(应用纲要的标题，人读)、Description(应用纲要的描述)、Resource Templates(应用纲要所包含的资源模板)等元素来描述。

基于书目框架设计的家谱本体即是一个领域本体，如何在系统中得到应用和实施可以用书目框架应用纲要来定义。应用纲要由标准的编码语言编写，可被机器处理。系统读取应用纲要定义的规则自动生成基于家谱本体的对象数据。家谱的书目框架应用纲要以 JSON 格式来定义，限于篇幅，这里只截取“纲要定义”、一个“资源模板”和两个“属性模板”的定义代码。“资源模板”以 shlgen：Person 为例，“属性模板”以 shlgen：family 和 shlgen：given-Name 为例。第一个属性的范围是 shgen：Family 类，第二个属性的范围是 Literal，其中“"type"："resource",”这行代码表示属性 shlgen：family 的范围是另一个资源对象，“"valueTemplateRefs"：["bfp：Family",]”指明是哪种资源对象，"bfp：Family" 指的是另一个资源模板的 ID，这个资源模板所定义的类

(shlgen：Family)是属性 shlgen：family 的范围。代码如下：

```
{
  "Profile"：  {
    "id"："bfp：Genealogy",
    "title"：" 家谱 ",
    "description"：" 上海图书馆家谱应用
      纲要。",
    "date"："2014-07-27",
    "contact"："cjxia@libnet.sh.cn",
    "resourceTemplates"：[
      {
        "id"："bfp：Family",
        "resourceLabel"："Family",
        "resourceURI"："http://gen.library.
          sh.cn/vocab/Person",
        "propertyTemplates"：[
          {
            "propertyURI"："gen.library.sh.
            cn/vocab/family",
            "propertyLabel"：" 家族 ",
            "mandatory"："false",
            "repeatable"："true",
            "type"："resource",
            "valueConstraint"：{
              "valueTemplateRefs"：[
              "bfp：Family",
          ]
            }
          }
          "propertyTemplates"：[
          {
            "propertyURI"："gen.library.sh.
              cn/vocab/givenName",
            "propertyLabel"：" 名字 ",
            "mandatory"："true",
            "repeatable"："false",
            "type"："Literal",
            "valueConstraint"：{
              "valueLanguage"："CHT",
              "languageLabel"：" 繁体中文 ",
            }
          }
          {  ...  }更多属性模板
        }
      {…}更多资源模板
    ]
  }
}
```

3.2 家谱本体的扩展

上述家谱本体中类和属性主要基于目前上海图书馆家谱数据的现状来设计，能够容纳现有家谱数据中的数据项。随着标引方法和技术的进步，如基于图像的标引技术，家谱数据中更多的数据项将在未来的标引工作中提取出来，比如家谱的世系图录包含家族中详细的成员名单和他们之间的亲属关系。目前上海图书馆的家谱世系图录只是扫描后作为图片存储，没有对图中的文字进行 OCR 识别，这部分内容是家谱资源中宝贵的财富，如果将来做更细粒度的标引，那么目前的本体就不够用，需要进一步扩展。一般来说，本体的扩展有复用已有本体和自定义本体两种做法，本文设计的家谱本体在模型和框架层面能够支持这两种做法。以世系图录为例，可以采用复用已有本体的办法。比如要复用 genOnt 来描述人与人之间的关系，可以为 shlgen：Person 类增加属于 genOnt 本体的属性，例如用 genont：hasFather、genont：hasMather 来表示父母子女的关系。如果还不够用，还可以关系本体(Relationship)的属性 rel：friendOf 表示朋友关系，域和范围均为 shlgen：Person。以人的墓志铭为例，目前已有的家谱数据中没有墓志铭的数据，但将来如果对《中国家谱资料选编》做标引，就需要对墓志铭作出定义。可以采用自定义新的属性来扩展目前的家谱本体，为shlgen：Person 增加一个属性 shlgen：epitaph，其域为shlgen：

Person，范围为文本串(Literal)。至于在本体扩展时究竟采用哪种方法，原则是尽量复用已有的较为成熟和被业界公认的本体，如果没有可复用的本体才考虑自定义。家谱本体扩展的目的是为数据实体增加相关的描述，使数据间的关系更丰富。由于数据的编码采用 RDF 数据模型，因而只需要增加一个或多个三元组，不影响后台数据的存储结构。

3.3 家谱本体的重用

知识本体是领域共享的知识，得到更多应用系统的重用才能体现更大的价值。本体的重用需要做好两方面的准备：一方面要准备供人读的翔实的说明文档，对类和属性的定义要明确，尽量避免在被重用的过程中产生歧义；另一方面要在 Web 上发布机器可读的基于标准编码语言的文档，一般用 RDFs 或 OWL 语言，在文档中声明前缀和命名空间，用规范的元素描述类和属性的定义。书目框架采用了 RDFs 的 9 个元素来对其本体词表编码，见表 2。

表 2 上海图书馆家谱本体的 RDFs 编码规则

标签	说明
rdfs：Class	定义一个类
rdfs：label	定义类或属性的标签
rdfs：subClassOf	定义一个类是另一个类的子类
rdfs：comment	定义类或属性的说明
rdf：about	定义类或属性的 URI
rdf：resource	定义资源的 URI
rdf：Property	定义一个属性
rdfs：domain	定义属性用于描述的类
rdfs：range	定义属性的取值范围

上海图书馆家谱本体也采用 RDFs 来定义。以下示例是对 bf：Work 类和自定义类 shlgen：Family 的定义，以 RDF/XML 格式编码：

```
<rdfs :Classrdf :about="http ://bibframe.org/
  vocab/Work">
<rdfs：label>Work</rdfs：label>
<rdfs ： subClassOfrdf :resource="http ://
  bibframe.org/vocab/Resource"/>
<rdfs：comment>Resource reflecting a con-
  ceptual essence of the cataloging resource.
</rdfs：comment>
</rdfs：Class>
<rdfs： Classrdf： about="http ://gen.library.sh.
  cn/vocab/Family">
<rdfs：label> 家族 </rdfs：label>
<rdfs :subClassOfrdf :resource="http ://
  bibframe.org/vocab/Family"/>
<rdfs：comment> 家谱相关的家族。</rdfs：
comment>
</rdfs：Class>
```

以下示例是对 bf：workTitle 属性和自定义属性 shlgen：ancestor 的定义，以 RDF/XML 格式编码：

```
<rdf： Propertyrdf： about="http： //bibframe.
  org/vocab/workTitle">
<rdfs :domainrdf :resource="http ://bibframe.
  org/vocab/Work"/>
<rdfs：label>Work title</rdfs：label>
<rdfs：rangerdf：resource="http：//bibframe.
  org/vocab/Title"/>
<rdfs：comment>Title or form of title chosen
  to identify the work，such as a preferred
  title，preferred title with additions，uniform
  title，etc..</rdfs：comment>
</rdf：Property>
<rdf：Propertyrdf：about="http：//gen.library.
  sh.cn/vocab/ancestor">
<rdfs：label> 始祖 </rdfs：label>
<rdfs：domainrdf：resource="http ://gen.library.
  sh.cn/vocab/Family"/>
<rdfs :rangerdf： resource="http：//gen.library.
  sh.cn/vocab/Person"/>
<rdfs：comment> 家族的始祖。</rdfs：com-
  ment>
</rdf：Property>
```

4 总结与展望

书目框架作为基于关联数据技术的本体模型，既能揭示家谱资源的文献特征，又能揭示其内容特征，并在家谱各种数据实体之间建立能被机器处理和理解的关联关系。这些措施能有效提高家谱系统的查全率和查准率，提升家谱资源服务的效果。

然而目前书目框架项目尚未结束，其模型仍在发展变化之中，一些细节尚未决定或仍在讨论和征求意见的阶段，这导致基于书目框架来设计家谱本体存在一定的风险性。因此，在家谱本体的设计过程中，主要以书目框架的核心模型和总体框架为基础模型框架，尽量避免复用存在争议或概念尚不明晰的类和属性，同时考虑架构的灵活性和可扩展性(好在基于关联数据的模型本身就具有这方面的优势)，以便今后进一步修订。

家谱本体设计的难点在于对人、地、时、事之间复杂关系的处理，尤其是家谱数据中对时间和地点的描述：不同时间同一地点的名称不一致、不同地点重名、同一地点在不同的时间范围内属于不同的行政区域划分、同一时间使用不同的纪年方式、时间范围的起止定位等问题为数据的清洗和实体对象的提取带来了困难。处理这些问题，需要引入已有的外部本体和规范词表，比如事件本体(Event Ontology)、时间本体(Time Ontology)、关系本体(Relationship Ontology)，以及即将以关联数据发布的Getty 的地理名词叙词表[30]等，来处理人、地、时、事之间的复杂关系，以补充现有家谱本体的不足。

下一步的工作是将以 RDFs 编码的家谱本体发布成关联数据，使之在 Web 上可访问可获取，可被其他本体复用，并提供数据消费接口(如SPARQL 端点)等，以达到方便地共享和重用的目的。同时，基于书目框架应用纲要开发应用系统，生成包含丰富关联的家谱对象数据，在这个过程中进一步检测家谱本体的健壮性和可靠性。

在我国，关联数据的介绍和试验已经有四五年，然而到目前为止，较大规模的实际应用还付之阙如。国外图书馆界最常见的关联数据应用是将国家书目库发布成关联数据，通常只有国家图书馆的数据才具有足够的规范性和权威性。选择家谱资源进行尝试，并采用书目框架作为本体模型，主要是基于上海图书馆家谱文献在质和量等方面于业界具有举足轻重的地位；同时，家谱资源无论多么特殊，都是上海图书馆馆藏文献的一部分，它需要遵从图书馆信息系统功能需求的一般性原则。

以关联数据为代表的语义技术对图书情报领域有着极为特殊的意义。上海图书馆正努力把该项目做成关联数据应用的一个示范性项目，希望能以此带动数字图书馆的资源揭示从基于文献向基于内容进行升级，为打造数字人文服务和研究平台进行具有突破意义的探索和尝试。

参考文献

[1] 刘炜，夏翠娟. 书目数据新格式 BIBFRAME 及其应用[J]. 大学图书馆学报，2014（5）：5-13.

[2] Tim Berners-Lee. Linked Data [EB/OL]. [2011-05-15]. http：//www.w3.org/DesignIssues /LinkedData.html.

[3] 刘炜. 关联数据：概念、技术及应用展望[J]. 大学图书馆学报，2011（2）：5-12.

[4] 王昭. 家谱文献资源整理现状与思考[J]. 中国科技信息，2013（5）：62-66.

[5] 毛建军. 中国家谱数字化资源的开发与建设[J]. 档案与建设，2007（1）：22-24.

[6] Campanya Artes Joan. The Family History Department of The Church of Jesus Christ of Latter-day Saints（LDS Church）. The GEDCOM Standard Release 5.5 Introduction[EB/OL]. [2014-05-11]. http：//home-pages.rootsweb.ancestry.com/~pmcbride/gedcom/55gcint.htm#S1.

[7] GENTECH Genealogical Data Model： A Comprehensive Data Modelfor Genealogical Research and Analysis（ version 1.1） [EB/OL].（2000-05-29）[2014-07-03]. https://www.ngsgenealogy.org/ngs-gentech/projects/Gdm/Gdm.htm.

[8] Jay Askren. The Semantic Web for Family History [EB/OL]. [2014-05-16]. http://jay.askren.net/Projects/SemWeb/

[9] Albert Meroño-Peñuela. Semantic Technologies for Historical Research: A Survey[EB/OL] .[2014-07-15]. http://www.semantic-web-journal.net/system/files/swj588.pdf.

[10] Josh Hansen. The Coming Web of Genealogical Data [EB/OL]. [2014-05-12]. http: //fht.byu.edu/prev_workshops/workshop12/papers/3.1%20Josh%20Hansen% 20-% 20FHT% 202012% 20Workshop% 20Paper% 20-%20The%20Coming%20Web%20of%20Genealogical%20Data.pdf.

[11] John Goodwin .John Goodwin's Family Tree[EB/OL]. [2014-07-08]. http://datahub.io/dataset/john-goodwins-family-tree.

[12] 周秋芳，顾燕，陈建华，等. 我国数字图书馆标准规范建设：家谱描述元数据规范[EB/OL].[2014-05-08].http：//www.docin.com/p-9321300.html.

[13] 赵亮，苏品红.国家数字图书馆工程标准规范成果：国家图书馆家谱元数据规范与著录规则[M].北京：国家图书馆出版社，2014：10-40.

[14] 上海图书馆. 中国家谱总目[M]. 上海：上海古籍出版社，2008：10-12.

[15] Ivo Zandhuis. Towards a Genealogical Ontology for the Semantic Web [EB/OL]. [2014-06-09]. http: // www.zandhuis.nl/sw/genealogy/.

[16] Charla Woodbury，David W. Embley. Family History Research on the Semantic Web：Building a Semantic Prototype for Danish Research[EB/OL].[2014-07-28]. http://fht.byu.edu/prev_workshops/workshop05/FHTCD/session1/s1-CharlaWoodbury_SemanticWeb.pdf.

[17] 陈艳.中国家谱的知识本体构建 [D]. 上海：上海交通大学，2007.

[18] Ying Jiang， Hui Dong. Ontology Based Knowledge Modeling of Chinese Genealogical Record[C]//Semantic Computing and Systems， 2008. WSCS '08. IEEE International Workshop：33-34.

[19] Dan Brickley，Libby MillerFOAF Vocabulary Specification 0.99 [EB/OL]. [2014-07-04]. http: //xmlns.com/foaf/spec/.

[20] Ian Davis， Eric Vitiello Jr. RELATIONSHIP: A vocabulary for describing relationships between people [EB/OL]. [2014-07-05]. http: //vocab.org/relationship/.html.

[21] GeoNames Team. GeoNames Ontology [EB/OL]. [2014-07-05]. http://www.geonames.org/ontology/documentation.html.

[22] Jerry R. Hobbs， Feng Pan. Time Ontology in OWL [EB/OL]. (2006-09-27) [2014-07-05]. http://www.w3.org/TR/owl-time/.

[23] Yves Raimond， Samer Abdallah The Event Ontology [EB/OL].[2014-07-09]. http: //motools.sourceforge.net/event/event.html.

[24] PeroniSilvio，TomasiFrancesca，VitaliFabio. Reflecting on the Europeana Data Model[M]. Digital Libraries & Archives，2013：228-240.

[25] Ted Fons， Jeff Penka， Richard Wallis. OCLC's Linked Data Initiative：Using Schema.org to Make Library Data Relevant On The Web [EB/OL].[2014-06-12]. http://www.niso.org/apps/group_public/download.php/9408/IP_Fons-etal_OCLC_isqv24no2-3.pdf.

[26] Library of Congress. Bibliographic Framework as a Web of Data：Linked Data Model and Supporting Services [EB/OL]. (2012-11-21) [2013-09-12]. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.

[27] 刘炜，李大玲，夏翠娟. 元数据与知识本体[J]. 图书馆杂志，2004 (6)：50：54.

[28] 叶鹰，金更达. 基于元数据的信息组织与基于本体论的知识组织[J]. 中国图书馆学报，2004 (4)：43-47.

[29] Library of Congress. BIBFRAME Profiles： Introduction and Specification [EB/OL]. [2014-06-18].http：//www.loc.gov/bibframe/docs/bibframe-profiles.html.

[30] The J. Paul Getty Trust. Getty Thesaurus of Geographic Names Online [EB/OL]. [2014-07-18]. http:// www.getty.edu/research/tools/vocabularies/tgn/index.html.

作者简介 夏翠娟，女，上海图书馆系统网络中心研究开发部高级工程师；刘炜，男，博士，研究员，上海图书馆副馆长；张磊，男，上海图书馆系统网络中心研究开发部高级工程师；朱雯晶，男，上海图书馆系统网络中心研究开发部工程师。

收稿日期 2014-08-14