

# 关联数据 URI 设计规范探讨\*

许磊 夏翠娟 刘炜 张磊

**摘要** 统一资源标识符(URI)是互联网领域重要的基础技术之一,是在网络虚拟空间标识和定位事物的基本方式。作为统一资源定位符(URL)的上位概念,URI在形式和功能上规范了互联网资源标识体系,为基于概念和模拟真实世界的关联数据应用提供了基本的标识和解析技术。当前业界关联数据应用正处于起步阶段,亟需引介推广重要的技术规范并交流相关实践。本文遵照万维网协会(W3C)有关URI设计原则的推荐规范,从URI的类型、设计原则和模式等三个方面介绍国外关联数据URI设计的最佳实践。最后结合上海图书馆的实践,设计关联数据应用的URI应用规范,以供同行参考。图2。表8。参考文献21。

**关键词** URI 设计规范 统一标识符 关联数据

## URI Design Patterns in the Context of Linked Data

Xu Lei Xia Cuijuan Liu Wei Zhang Lei

**Abstract:** As one of the basic technologies of the semantic web, Uniform Resource Identifier (URI) is a crucial approach to identify resources on the internet. Different from the common Uniform Resource Locator (URL), URI is a generic concept to regulate the internet resource identification system and provide basic identification and resolving methods for linked data technology. The linked data applications have just started in its infancy. So it is necessary to promote the implementation of related technical specifications and to exchange relevant practices. Three aspects of URI design best practices in abroad, including URI types, design principles and patterns, are introduced in accordance with the related URI design principles of World Wide Web Consortium (W3C). Finally, this paper gives an explicit explanation as the best practice of URI guidelines for the linked open data applications in Shanghai Library. 2 figs. 8 tabs. 21 refs.

**Keywords:** URI Patterns; Unique Identifier; Linked Data

### 1 引言

名称通常是事物最重要的属性。古希腊哲学家巴门尼德认为,事物没有名称就可以被认为是不存在的。互联网上的资源正是如此。万维网协会(World Wide Web Consortium, W3C)将网络资源定义为:“任何用URI标识的对象”<sup>[1]</sup>,这就是说网络上任何资源,其唯一且必需的属性是URI,即统一资源标识符(Uniform Resource Identifier)。有了URI的资源就是网络“公民”,就能够被管理、存储、跟踪、调用,进行生命周期管理;没有

URI的资源就不是网上的一种独立存在,必须附属於其它具有URI的资源。网页是万维网上的一种常见资源,其URI就是URL(Uniform Resource Locator,统一资源定位符)。万维网发明人蒂姆·伯纳斯-李于2006年提出“关联数据”概念,推动万维网由文献的Web向数据的Web发展,关联数据四原则中第一条即“使用URI命名事物”<sup>[2]</sup>。

URI是一组按照特定语法规则构成的字符串,为网络资源的标识提供了一种简单和可扩展的方法<sup>[3]</sup>。URI作为网络中对象的唯一标识符,

\* 本文系国家哲学社会科学基金青年项目“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)的研究成果之一。

经其标识的地址可以获得关于被标识对象的有用信息,并可关联到其他相关资源。URI 的特点体现在“统一”“资源”和“标识符”三个方面<sup>[3,4]</sup>。“统一”保证了不同类型的资源标识符可以采用统一的使用方式,并为新的资源类型提供统一的标识框架;也保证了这些不同类型的资源标识符在相同的语法规范下有统一的语义解释,并允许在不同的环境下重用。“资源”是指能够被标识的任何对象,不只限于网络文档和数字内容,也包括现实世界的物理实体和抽象概念。“标识符”是一个对象可以引用的名称。URI 使对象能够实现标识和定位、访问与获取、参引和规范控制等目标。

URI 设计是关联数据应用过程中至关重要的一环,合理的 URI 设计模式,是数据得到更好地组织和管理、传输和交换、共享和重用的保障。

## 2 URI 的作用

### 2.1 标识和定位

URI 的本质是一种标识符。所谓标识符,是指对象的名称标识,是特定背景下资源的明确指向。标识符的基本功能就是对资源进行标识,即对资源的命名和区分。如 ISBN 用 7501309086 与 7544130967 标识了两本同名但不同内容的书。在一个开放式的网络世界中,使用全球唯一的 URI 对实体对象进行标识,可以有效地对不同名称的相同对象,以及相同名称的不同对象进行命名和区分。

URI 可以进一步地分为统一资源定位符(URL)、统一名称(URN)或两者的结合<sup>[3]</sup>。ISBN 号就只是一个 URN,而 telnet://192.0.2.16:80/ 则是一个 URL。在语义网语境下,URI 可被视为定位符,不仅可以标识一个资源,而且能够提供该资源的定位信息。

### 2.2 访问与获取

严格意义上,URI 本身并不总是提供资源的访问信息,它只对资源标识提供保障。同样地,在语义网语境下,URI 同时作为统一名称和定位符

而存在,它不仅命名了该资源,还提供了定位该资源的主要访问机制。如 <http://www.ietf.org/rfc/rfc3986.txt>,不仅标识定位 RFC3986 这份文档资源,也可以通过 HTTP 协议进行访问并获取相关内容。

### 2.3 参引

参引是指为了获取引用资源的相关信息,在万维网上查找 URI 的过程<sup>[5]</sup>。可参引的 URI 提供了使用任何网络传输协议获取被标识资源的表示形式的信息检索机制。如,传统网页检索中,对 URL 的参引可以直接返回 HTML 页面。而在关联数据环境中,对 URI 的参引可以返回 HTML、XML 或 JSON 等形式。因此,参引一个 URI,将指向一个资源的描述,这个描述中可能会有更多的 URI 链接,同时这些 URI 也是可参引的,最终个体资源互相关联形成一个数据网。

### 2.4 规范控制

规范控制是图书馆编目工作中的重要一环,通过建立规范检索点保障书目系统中重要属性描述的一致性,满足用户查找、识别、提供情境、证明、选择和探索等需求。传统的规范控制工作是运用信息资源标目法,确定人名、团体、地名、题名、主题等检索点及其规范形式,同时建立和维护规范文档的过程<sup>[6]</sup>。但随着信息爆炸和载体类型的多样化,严格的规范控制工作对于人工来说已不堪承受。

规范控制的本质是实现基于概念的描述和匹配,即利用表征概念体系的符号体系构建一个概念空间。而以 RDF 为基础的关联数据技术,其目的正是构建概念化的知识空间<sup>[7]</sup>。URI 标识的实体概念将独立于其表示形式,各种语言或符号的表达都可以作为标签属性。URI 就是传统的标目,任何形式的符号在检索和显示中都将具有相同的优先级。而概念之间的语义关系,也可以通过 RDFS、SKOS、OWL 等编码模式得到清晰的描述。另外,URI 可以标识任何对象,也就是说书目数据中的任何属性都是可控制的,传统意义上的规范数据将得到极大的丰富。在元数据生产过程中,规范控制也将无处不在。

### 3 URI 设计与应用案例分析

在关联数据实践中,为每一个对象分配唯一标识符 URI 至关重要。URI 确保了被标识对象的识别、定位、参引、重用和规范控制。因此,设计一个“酷”的 URI 成为必须。“酷”URI 不仅为人机提供描述,同时也应该简洁、稳定和易于管理<sup>[8]</sup>。为此,政府部门、图书馆和 W3C 等机构团体都对 URI 设计原则和模式进行了研究。

#### 3.1 URI 标识资源类型

URI 标识的对象不仅有传统的网络资源,如网页、图片、视频等;也包括人、团体、作品、时间、概念等非网络资源,相对于网络信息资源,这些“真实世界的事物”统称为非信息资源<sup>[9]</sup>。由于标识对象类型的不同,因而在后续的设计模式与技术实现环节会有不同的处理方式。因此,在制定 URI 设计原则时,首先需要对 URI 类型进行区分。

按照被标识对象的类型,URI 可以简单分为信息资源 URI (Resource URI) 和非信息资源 URI (Thing URI)。但在实践过程中,由于数据类型的多样性,URI 类型会进一步地细分。Sonya Abbas 和 Adegboyega Ojo 就综合已有的 URI 设计实践将 URI 分为了 6 种类型:事物 URI (Thing URI),概念 URI (Concept URI),资源 URI (Resource URI),表达 URI (Representation URI),分层 URI (Hierarchical URI) 和本体 URI (Onto URI)<sup>[10]</sup>。其中,概念 URI 标识的对象主要是抽象的概念、主题;表达 URI 是与内容协商机制有关的 URI,具体内容将在 3.3 中加以说明;分层 URI 更多的是出于结构角度而非标识对象类型角度分类,表示了 URI 的层次结构。

具体到图书馆领域,W3C 图书馆关联数据孵化小组根据数据的使用特性将图书馆数据分为三类:数据集、元素集和属性值词汇。数据集是结构化元数据的集合,用来描述实体,主要即为书目数据集。元素集定义了描述实体的类和属性,如 FRBR、BIBFRAME 等。属性值词汇定义元数据记录中资源(如主题、艺术类型、作者)元素的

属性值范围,如 LCSH、VIAF、DDC 等<sup>[11]</sup>。

#### 3.2 政府及机构 URI 设计原则与应用

作为关联数据最主要的推动者,W3C 在 2008 年即发布了有关 URI 实施的指导性文件 *Cool URIs for the Semantic Web*<sup>[8]</sup>。2011 年,W3C 又发布了政府关联数据 URI 设计的最佳实践,提出了 3 条基本原则:使用 HTTP URIs,提供至少一种机读格式和不使用变量字符串,以及 URI 稳定性和国际化的要求<sup>[12]</sup>。

2007 年,欧盟空间信息基础设施建设 (INSPIRE) 2007/2/EC 指令提出“拥有通用结构的对象标识符在各国系统间的映射可以确保互操作性”<sup>[13]</sup>,从政策方面阐述了 URI 设计原则和模式的必要性,推动欧盟成员国制定国家层面的设计原则。2012 年“欧洲公共机构互操作性解决方案”(Interoperability Solutions for European Public Administrations, ISA) 发布了由 Phil Archer 等人编写的关于永久性 URI 的研究报告<sup>[14]</sup>。该报告从 URI 格式、设计原则和管理三个方面对欧盟机构及其成员国、国际性组织和文化遗产领域进行了调研。在此基础上,Phil Archer 提出了 URI 的建议模式和原则。其 URI 模式为 `http://|domain|/|type|/|concept|/|reference|`,基地址 |domain| 由服务器地址和政府类型组成,如 `education.data.gov.uk`; |type| 是表明被标识对象类型的词汇,如 `id` 表示非信息资源,`doc` 表示信息资源,`def` 表示本体概念,`set` 表示数据集; |concept| 表示具体对象集合,如 `person`、`road`、`school` 等; |reference| 表示标识符。URI 设计原则共 10 条,具体内容见表 1。

表 1 欧盟 ISA 的 URI 设计原则<sup>[15]</sup>

URI 设计推荐做法	URI 设计应避免的做法
遵循特定的 URI 模式	避免出现版权说明
复用已有的标识符	避免出现版本说明
内容协商	避免使用自增量
303 重定向	避免使用查询字符
使用专用的服务	避免使用文件扩展名

2009年,英国发布了政府部门URI设计原则 *Designing URI Sets for the UK Public Sector*,并于2014年发布新版的 *URI Patterns*<sup>[15]</sup>。英国政府关联数据的URI类型除有非信息资源URI、信息资源URI、规范词表URI(包含规范数据、取值词表和本体)和数据集URI外,还定义了URI集

URI(URI set URI)。它提供了关于URI集合的管理元数据信息,如URI集的时空信息、主题范围、出处、数据质量等。在其他机构复用URI时,URI集URI能够提供背景信息,保证数据的权威性和可追溯性。英国政府各类型URI的模式见表2。

表2 英国政府开放数据URI模式<sup>[15]</sup>

URI type	URI pattern	备注
Set URI	{ prefix } / id / { concept } or { prefix } / { concept } # id	{ prefix } = http:// { domain } / { collection } *
Identifier URI	{ prefix } / id [ / { concept } / { key } ] * or { prefix } [ / { concept } / { key } ] * # id	{ domain } = { sector } . data . gov . uk  / { collection } * } = 有共同内涵的分组 ( URI set , datasets , vocabulary ) 。主要目的就是保证即使组织变更 ( domain ) 也可以维持 / collection / 的稳定
Document URI	reference data for single reference items ; { prefix } / doc [ / { concept } / { key } ] * or { prefix } [ / { concept } / { key } ] * optionally , reference data for lists of reference items ; { prefix } / doc / { concept } / { key } ] * / { concept } or { prefix } [ / { concept } / { key } ] * / { concept }	{ type } = def , id , doc , data , so ( 空间 )  { concept } = 如 person , road , school 等概念词汇表示 URI 标识对象的本质
Vocabulary URI ( for vocabularies , ontologies , concept schemes , codelists and schema )	{ prefix } / def / vocabulary * }	{ key } = 标识符 / { vocabulary } * } = 术语集 , 如 schema , codelist , concept scheme or ontology
Vocabulary Term URI ( for term definitions within a vocabularies , ontologies , concept schemes , codelists and schema )	{ prefix } / def / vocabulary * } / { term }	
Dataset URI ( for datasets )	{ prefix } / data / dataset * }	
Data Item URI ( for data items within datasets )	{ prefix } / data / dataset * } [ / { concept } / { key } ] * }	

在参考英国政府2009年的实践后,澳大利亚政府于2014年发布了自己的URI设计原则<sup>[16]</sup>。澳大利亚政府关联数据将URI分为数据集URI、非信息资源URI、信息资源URI和本体URI等4种类型,同时从数据集URIs、域名结构、URI模式、URIs发布和URI命名约定等5个角度规定了多达40条的原则,并将其分为必须、不可以、应该、不应该、建议、可以和可选各种等级。其中

“必须”原则就有12条,如必须使用HTTP URI,至少提供一种机读格式和人工可读的HTML,URI的基地址是data.gov.au等。同英国的URI集URI类似,澳大利亚在原则5—7中,以“应该”等级规定了为数据集URI提供出处、质量等元数据信息,并“建议”使用数据目录词表(DCAT)。澳大利亚政府关联数据建议的URI模式见表3。

表3 澳大利亚政府开放数据的 URI 模式<sup>[16]</sup>

URI 类型	URI 模式
Dataset URI	{ domain }. data. gov. au/dataset/[ / { module } ] * / { datasetid }
Identifier URI	{ domain }. data. gov. au/dataset/[ / { module } ] * /id/ { type } / { name }
Document URI	{ domain }. data. gov. au/dataset/[ / { module } ] * /doc/ { type } / { name }
Ontology	{ domain }. data. gov. au/dataset/[ / { module } ] * /def#

注:[ ]表示可选,\*表示数量不定,module表示数据集细分类型,domain表示政府职能,取值词表为澳大利亚政府互动功能叙词表(AGIFT)。

### 3.3 图书馆 URI 设计原则与应用

在图书馆领域,虽没有发布单独的 URI 设计文档,但在 URI 设计中都遵循了 Cool URI 的基本原则和统一的模式。2011年,大英图书馆将其书目数据发布为关联数据,其标识的对象有人、团体、家族、主题、事件、书目等。英国国家书目数据

(The British National Bibliography, BNB)的 URI 按照非信息资源、信息资源、规范词表和本体分为4类<sup>[17]</sup>。在 BNB 的 URI 设计中,将人、团体、家族和地点同时作为非信息资源和主题概念进行处理,两者的 URI 采用不同的模式,并使用 foaf: focus 进行关联,具体内容见表4。

表4 大英图书馆 URI 模式

URI 类型	URI 模式	
人	Person (as Agent)	http://bnb. data. bl. uk/id/person/ { person-name }
	Person (as Concept)	http://bnb. data. bl. uk/id/concept/person/lcsh/ { person-name }
	Birth	http://bnb. data. bl. uk/id/ { person-name } /birth
	Death	http://bnb. data. bl. uk/id/ { person-name } /death
团体	Organization (as Agent)	http://bnb. data. bl. uk/id/organization/ { organization-name }
	Organization (as Concept)	http://bnb. data. bl. uk/id/concept/organization/lcsh/ { organization-name }
家族	Family (as Agent)	http://bnb. data. bl. uk/id/family/ { family-name }
	Family (as Concept)	http://bnb. data. bl. uk/id/concept/family/lcsh/ { family-name }
地点	Place (as “Thing in the World”)	http://bnb. data. bl. uk/id/place/ { place-name }
	Place (as Concept)	http://bnb. data. bl. uk/id/concept/place/lcsh/ { place-name }
主题	Topic (LCSH)	http://bnb. data. bl. uk/id/concept/lcsh/ { topic }
	Topic (Dewey)	http://bnb. data. bl. uk/id/concept/ddc/ { edition-number } / { dewey-number }
出版	Publication event	http://bnb. data. bl. uk/id/resource/ { control - number } /publicationevent/ { publicationeventstring }
	Publisher/Distributor	http://bnb. data. bl. uk/id/agent/ { name-publisher/distributor }
文献	Series	http://bnb. data. bl. uk/id/series/ { series-name+ISSNifpresent }
	Bibliographic resource	http://bnb. data. bl. uk/id/resource/ { control-number } http://bnb. data. bl. uk/id/resource/ { bnb-number }

美国国会图书馆(Library of Congress, LC)于2009年将其规范数据发布为关联数据,截至目前,其关联数据词表已经包括规范部分、保存词汇表、MARC代码和取值词表与代码表4大类共54种<sup>[18]</sup>。LC的URI类型分为规范数据URI、取值词表URI以及本体URI。URI模式是http://id.loc.gov/{type}/{concept}/{reference},其中type的取值有authorities、vocabulary和ontologies, Concept取值有subjects、relators、preservation、roles等。

法国国家图书馆(Bibliothèque Nationale De

France, BNF)在2011年将责任者、主题和作品发布为关联数据。BNF的URI类型分为了非信息资源URI(包含规范数据URI)、本体URI和取值词表URI<sup>[19]</sup>。非信息资源URI模式为http://data.bnf.fr/ark:/12148/{reference},本体URI是http://data.bnf.fr/ontology/bnf-onto/,取值词表URI是http://data.bnf.fr/vocabulary/{type}/{reference}。其他图书馆,如德国国家图书馆、日本国会图书馆的URI设计模式基本相同<sup>[20]</sup>。表5对部分图书馆关联数据项目的URI类型进行了统计。

表5 图书馆关联数据项目URI类型统计

机构	非信息资源URI (entity URI)	信息资源URI (resource URI)	规范词表URI (authority URI)	取值词表URI (vocabulary URI)	本体URI (ontology URI)
大英图书馆	√	√	√		√
美国国会图书馆	√		√	√	√
OCLC	√		√		
法国国家图书馆	√		√	√	√
德国国家图书馆	√	√	√	√	√
瑞典国家图书馆	√	√			
西班牙国家图书馆	√	√	√		√
匈牙利国家图书馆	√	√	√		
日本国会图书馆	√		√		√
芬兰国家图书馆			√		√
意大利众议院图书馆	√	√	√		√
奥斯陆公共图书馆	√	√	√	√	√

注:法国、德国和西班牙国家图书馆的非信息资源URI、规范词表URI或取值词表URI并没有严格按照表中的类型进行区分。

#### 4 上海图书馆的URI设计原则

参考已有的设计原则和Sonya Abbas、Adegboyega Ojo的URI设计模式,上海图书馆确定其URI设计原则为:无变量、稳定性、使用HTTP URI、可读性和国际化。

##### 4.1 无变量

无变量即为保持URI的稳定和可靠,在URI字符串中应避免使用文件扩展名、所有权信息、

版本信息等变量内容,避免包含主题词、自动增量,避免与具体时间的技术捆绑。如http://example.com/people/v01/123456、http://example.com/people/123456.rdf就不是“酷”的URI。

URI一般由三部分组成:访问资源的命名机制,存放资源的主机名和资源自身的名称或路径。接下来就是两个可选部分,以“?”开头的查询字符串和以“#”开头的片段标识。其中,查询部分是没有结构的字符串,其语法、语义取决于具体的

应用,没有统一规定。因此,此部分不应出现在 URI 的构成中,即避免与具体的技术捆绑。

### 4.2 稳定性

研究发现学术出版物的 URL 在发布 7—10 年后,有一半会失效<sup>[21]</sup>。图书馆作为人类文化知识的保存和传播中心,在发布数据时,应该为标识对象分配稳定长久的 URI,以确保资源长久地可获取和重用。因此在 URI 设计中应遵循一致的表达、统一的模式,并复用已有的内部标识符。如美国国会图书馆的 URI 统一模式为 http://id.loc.gov/{type}/{concept}/{内部标识符},“鲁迅”的 URI 是 http://id.loc.gov/authorities/names/n50047988,n50047988 即为 LC/NACO 的内部标识符。

### 4.3 使用 HTTP URI

URI 语法<sup>[3]</sup>规定其以表示资源访问协议的 [schema:] 开头,常见的有 http、ftp、mailto、file 等。依据关联数据四原则<sup>[2]</sup>,选用 HTTP 作为 URI 的访问协议。HTTP 作为主流的网络传输协议,保证了最广泛的工具和系统可以访问。任何资源均可由 URI 标识。因此,当客户端参引 HTTP URI 时,需要根据所标识对象的类型和表示形式支持 303 转向和内容协商机制。

#### 4.3.1 支持 303 转向

由上文已知,URI 标识的对象可简单地分为非信息资源和信息资源。当参引信息资源 URI 时,服务器端直接返回请求的网页。而非信息资源 URI 不可以被直接参引,服务器会返回 303 See Other 的状态码给客户端,表示“这是一个非信息资源,因此无法为您提供资源本身,但可以为您提

供一些链接,帮助您获得更多相关信息”。这个过程就是 303 重定向。

另外,非信息资源 URI 有两种形式,Hash URI 和 Slash URI。Hash URI 一般用于小型的相对稳定的数据集,在参引 Hash URI 时,HTTP 客户端通过剥离#后面的部分而返回整个 URI,以达到类似于 303 重定向的效果<sup>[5,8]</sup>;Slash 则适用于大型的、经常进行更新的数据集。澳大利亚政府 URI 文档中建议,当关联数据发布机构对域名、web 服务器拥有控制权,并且数据集中的实体数量较多时使用 Slash URI,否则使用 Hash URI<sup>[16]</sup>,如图 1 所示。

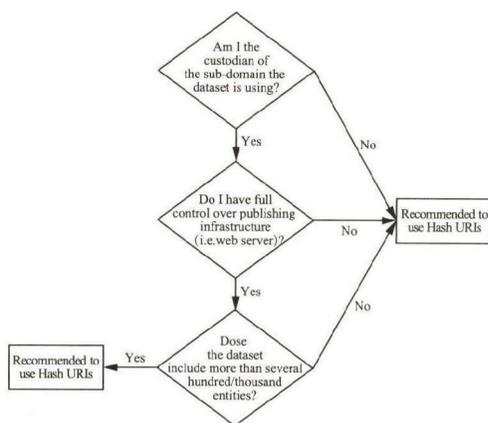


图 1 Hash URI 和 Slash URI 的选择流程<sup>[16]</sup>

#### 4.3.2 内容协商

一个 URI 标识的资源可以有不同的表示形式,如 HTML、RDF、PDF 等格式,或不同的语种。因此,当客户端参引 URI 时,服务器端需要根据其发送请求的 HTTP 报头返回合适的表示形式。图 2 显示了参引非信息资源 URI 时 303 重定向和内容协商的结合。

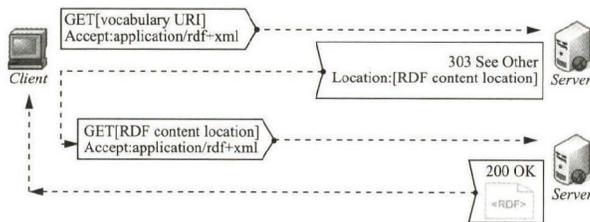


图 2 303 重定向与内容协商<sup>[5]</sup>

如参引法国国家图书馆“维克多·雨果”的 URI <http://data.bnf.fr/ark:/12148/cb11907966z>, 就会 303 重定向到 [http://data.bnf.fr/11907966/victor\\_hugo/](http://data.bnf.fr/11907966/victor_hugo/), 并根据 HTTP 报头进行内容协商返回合适的格式<sup>[19]</sup>:

[http://data.bnf.fr/11907966/victor\\_hugo/fr.html](http://data.bnf.fr/11907966/victor_hugo/fr.html)  
[http://data.bnf.fr/11907966/victor\\_hugo/en.html](http://data.bnf.fr/11907966/victor_hugo/en.html)  
[http://data.bnf.fr/11907966/victor\\_hugo/rdf.xml](http://data.bnf.fr/11907966/victor_hugo/rdf.xml)  
[http://data.bnf.fr/11907966/victor\\_hugo/rdf.n3](http://data.bnf.fr/11907966/victor_hugo/rdf.n3)

#### 4.4 可读性

URI 作为标识对象实体的一组字符串, 在保障唯一性和稳定性的前提下, 需要简短易记、利于人读、同时尽量使用小写字母(本体中的类名除外), 以便于管理和使用。

#### 4.5 国际化

URI 的使用是全网域的, 为了便于 URI 的复用和解析, 应使用 UTF8 编码, 不使用特殊字符, 如汉字、下划线等。

### 5 上海图书馆的 URI 设计模式

在设计 URI 模式时, 首先需要考虑本馆的数据类型, 据此进行 URI 类型的划分。再根据 URI 设计原则, 制定适合的 URI 模式。

上海图书馆关联数据项目包括了家谱数据、盛宣怀档案数据和书目数据, 以及本体、实体数据、规范数据和取值词表等。因此, 该馆将 URI 分为 6 种类型: 数据集 URI、本体 URI、规范词表 URI、取值词表 URI、非信息资源 URI 和信息资源 URI。数据集 URI 是用于特定资源的 URI, 如家谱数据集 URI; 规范词表 URI 主要标识对象为主题、分类等; 取值词表 URI 主要标识对象为责任

方式、版本类型、装订方式等取值词汇; 非信息资源 URI, 即标识人、地、时、团体等实体对象的 URI; 信息资源 URI 则标识作品、实例、单件等对象。

#### 5.1 总体设计模式

根据上文调研结果, 在 URI 设计模式中, 一般将要标识的对象分为本体词表、规范词表、非信息资源、信息资源、取值词表等类型。根据 URI 设计的 5 条原则, 上海图书馆的 URI 模式由域名 (domain)、要标识的对象所属的类型词 (type)、本体中的概念 (concept) 或取值词表中的术语所属的类型词 (termtyp) 、内部唯一标识符四个部分组成:  $http://\{ domain \}/\{ type \}/[ concept ] [ termtyp ] / \{ IRI \}$ 。各部分内容如下。

(1) domain = Base URI + [ dataset prefix ]。当所发布的数据为诸如机构、地名、时间等通用的规范数据时, dataset prefix 省略, 详见表 6。

(2) type 的取值详见表 7。

(3) concept 或 type。concept 取本体中的类名, type 为自定义的类型词。如, edition 表示版本取值词表, person 表示非信息资源的人, work 表示信息资源的作品。

(4) IRI, 系统唯一内部标识符。由于上海图书馆历史数据的规范控制不完善, 资源与实体没有统一的内部标识符。因此, 在新发布的关联数据中, 为每一个对象重新分配了系统唯一标识符。根据数据的不同, IRI 分为两种类型, 一种是标识本体、规范数据和取值词表, 另一种是信息资源和实体。前者使用术语的英文或拼音形式, 后者使用 16 位随机符号, 由 26 个小写字母和 1 至 9 数字的混合。这即遵循了无变量和国际性的 URI 设计原则, 也保证了 IRI 有足够的冗余量。

表 6 上海图书馆 URI 模式 domain 示例

类型	domain	备注
本体	<a href="http://www.library.sh.cn/ontology/">http://www.library.sh.cn/ontology/</a> 或 <a href="http://ontology.library.sh.cn/">http://ontology.library.sh.cn/</a>	同一机构所有本体使用相同 domain
关联数据	<a href="http://data.library.sh.cn/">http://data.library.sh.cn/</a>	
数据集	<a href="http://data.library.sh.cn/jp/">http://data.library.sh.cn/jp/</a>	家谱数据集
面向读者服务的网站	<a href="http://jp.library.sh.cn/">http://jp.library.sh.cn/</a> 或 <a href="http://jiapu.library.sh.cn/">http://jiapu.library.sh.cn/</a>	

表7 上海图书馆 URI 设计模式中的标识对象所属类型词(type)定义列表

标识对象类型词	说明
ontology	本体
authority	规范词表,如名称规范档、主题词、分类法等
entity	非信息资源:人、地、机构等实体
resource	信息资源:如图书、连续出版物等文献,以及文档、图像、视频、音频等数字化的资源对象。
vocab	取值词表,如纂修方式、先祖类别、版本类型

## 5.2 数据集的 URI 设计模式

数据集 URI 主要标识对象是特定数据集,如家谱数据集、名人档案数据集、书目数据集等。URI 模式为 `http://data.library.sh.cn/{dataset}/`。dataset 是数据集类型词汇,如家谱数据集 URI 为 `http://data.library.sh.cn/jp/`。

## 5.3 本体词表的 URI 设计模式

本体 URI 作为一种非信息资源 URI,有 Hash 和 Slash 两种形式。而上海图书馆的本体词表,包括家谱本体、书目本体、档案本体等,将共用一个域名。该本体词表将是一个不断更新和进化的大型词表。另外,作为该本体词表基础的 BIB-FRAME 使用了 Slash 形式。因此,上海图书馆的本体 URI 使用 Slash,模式为:`http://{domain}/ontology/[class][property]`。其中,Class 是首字母大写的类名,如 `http://www.library.sh.cn/ontology/Person`;Property 是首字母小写的属性名,如 `http://www.library.sh.cn/ontology/name`。

## 5.4 规范数据的 URI 设计模式

规范数据 URI 主要标识对象是主题和分类。URI 模式为:`http://{domain}/authority/{termtype}/{term}`。Type 是主题分类的类型,如 FDC 表示四部分类法,LGJ 表示刘国钧分类法。Term 是主题分类术语,如“jing”表示四部分类法中的“经”。完整的 URI 模式如:`http://data.library.sh.cn/authority/FDC/jing`。

## 5.5 信息资源的 URI 设计模式

信息资源 URI 主要标识对象是作品、实例、

单件等。URI 模式为:`http://{domain}/resource/{concept}/{IRI}`。其中,Concept 表示资源类型,如 work 表示作品,item 表示单件;IRI 为 16 位字母和数字随机组合的内部唯一标识符,如 `http://data.library.sh.cn/jp/resource/work/ullep7uwac4n8c8bt` 标识了题为“敖氏支修家谱”的家谱作品。

## 5.6 非信息资源的 URI 设计模式

非信息资源 URI 主要标识对象是人、地、时等。URI 模式与信息资源 URI 相同,如 `http://data.library.sh.cn/entity/person/mwe1ui6bww62a417` 标识了姓名为“敖开甫”的实体人。

## 5.7 取值词表的 URI 设计模式

取值词表的 URI 主要标识的对象是属性的取值,属性的取值范围限定在一个特定的词汇列表。根据取值词表的适用范围,上海图书馆将取值词表分为通用取值词表和专用取值词表。前者如版本类型的取值词表包括“石印本”“木活字本”“铅印本”等,装订方式的取值词表包括“经折装”“蝴蝶装”“散叶”等;后者如先祖类型为家谱专用取值词表,其取值词表包括“始祖”“始迁祖”等。

上海图书馆的取值词表 URI 模式为:`http://{domain}/vocab/{termtype}/{term}`。通用取值词表的 domain 为 BaseURI,专用取值词表为 BaseURI+dataset prefix。termtype 表示取值词表的类型,如 ancestor 表示先祖,edition 表示版本。term 为取值词的拼音全拼,中间用中划线分隔。

如,先祖类型的取值词“始迁祖”的 URI 为 <http://data.library.sh.cn/jp/vocab/ancestor/shi-qian-zu>, 装订方式的取值词“经折装”的 URI 为 [http://](http://data.library.sh.cn/vocab/binding/jing-zhe-zhuang)

[data.library.sh.cn/vocab/binding/jing-zhe-zhuang](http://data.library.sh.cn/vocab/binding/jing-zhe-zhuang).

表 8 列举了按照 URI 设计规范上海图书馆家谱关联数据中各类型资源 URI 的设计模式。

表 8 上海图书馆家谱 URI 设计模式

类型	设计模式	举例
本体	<a href="http://{domain}/ontology/{class}]{property}">http://{domain}/ontology/{class}]{property}</a>	<a href="http://www.library.sh.cn/ontology/">http://www.library.sh.cn/ontology/</a> 或: <a href="http://ontology.library.sh.cn/">http://ontology.library.sh.cn/</a>
规范数据	<a href="http://{domain}/authority/{termtype}/{term}">http://{domain}/authority/{termtype}/{term}</a>	<a href="http://data.library.sh.cn/authority/CCT/A">http://data.library.sh.cn/authority/CCT/A</a> (中国分类主题词表) <a href="http://data.library.sh.cn/authority/LGJ/000">http://data.library.sh.cn/authority/LGJ/000</a> (刘国钧分类法) <a href="http://data.library.sh.cn/authority/FDC/jing">http://data.library.sh.cn/authority/FDC/jing</a> (四部分类法) <a href="http://data.library.sh.cn/authority/familyname/ding">http://data.library.sh.cn/authority/familyname/ding</a> (姓氏)
实体	<a href="http://{domain}/entity/{concept}/{IRI}">http://{domain}/entity/{concept}/{IRI}</a>	<a href="http://data.library.sh.cn/jp/entity/person/6ad6f4b81">http://data.library.sh.cn/jp/entity/person/6ad6f4b81</a> (人) <a href="http://data.library.sh.cn/entity/organization/6ad6f4b81dadf">http://data.library.sh.cn/entity/organization/6ad6f4b81dadf</a> (机构) <a href="http://data.library.sh.cn/entity/place/dadfadfadfdd">http://data.library.sh.cn/entity/place/dadfadfadfdd</a> (地)
资源	<a href="http://{domain}/resource/{concept}/{IRI}">http://{domain}/resource/{concept}/{IRI}</a>	<a href="http://data.library.sh.cn/jp/resource/work/6ad6f4b81">http://data.library.sh.cn/jp/resource/work/6ad6f4b81</a> (创造性作品)
通用取值词表	<a href="http://{domain}/vocab/{termtype}/{term}">http://{domain}/vocab/{termtype}/{term}</a>	<a href="http://data.library.sh.cn/vocab/binding/jing-zhe-zhuang">http://data.library.sh.cn/vocab/binding/jing-zhe-zhuang</a> (经折装) <a href="http://data.library.sh.cn/vocab/edition/gao-ben">http://data.library.sh.cn/vocab/edition/gao-ben</a> (稿本)
家谱取值词表	<a href="http://{domain}/vocab/jp/{termtype}/{term}">http://{domain}/vocab/jp/{termtype}/{term}</a>	<a href="http://data.library.sh.cn/jp/vocab/ancestor/shi-qian-zu">http://data.library.sh.cn/jp/vocab/ancestor/shi-qian-zu</a> (始迁祖)

## 6 结语

随着关联数据在图书馆、档案馆等文化遗产机构的广泛应用,将会有更多的文化遗产资源与实体数据链接到关联开放数据的云图之中。URI 作为其基础技术之一,是标识世间万物的重要手段。永久稳定的 URI,保证了被标识对象的可访问和重用,这符合图书馆的目标和职责。因此,图书馆在实施关联数据项目时,可参考国内外关联数据应用中 URI 设计模式所遵循的总体框架和最佳实践,再根据实际需求制定适合本馆的 URI 设计模式,保证项目实施的一致性和稳定性,促

进资源的全网域共享和融合。

## 参考文献

- 1 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).
- 2 Berners-Lee, T. Linked Data [EB/OL]. [2016-02-22]. <http://www.w3.org/DesignIssues/LinkedData.html>.
- 3 Uniform Resource Identifier (URI): Generic Syntax [EB/OL]. [2016-02-22]. <http://www.ietf.org/rfc/rfc3986.txt>.
- 4 孙坦,等.国家图书馆数字资源唯一标识符规

- 范和应用指南[M].北京:国家图书馆出版社,2010:37.
- 5 Chris Bizer, et al. How to Publish Linked Data on the Web[EB/OL]. [2016-02-22]. <http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/#relationaldata>.
- 6 万爱雯,周建清.图书馆资源建设与编目工作研究[M].北京:当代中国出版社,2013:108.
- 7 刘炜,等.万维网时代的规范控制[J].中国图书馆学报,2015(3).
- 8 Leo Sauermann, Richard Cyganiak. Cool URIs for the Semantic Web[EB/OL]. [2016-02-22]. <http://www.w3.org/TR/cooluris/>.
- 9 Rhys Lewis. Dereferencing HTTP URIs[EB/OL]. [2015-09-03]. <http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14#sec-information-resources>.
- 10 Sonya Abbas, Adegboyega Ojo. Applying Design Patterns in URI Strategies - Naming in Linked Geospatial Data Infrastructure[J/OL]. [2015-09-03]. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6758863>.
- 11 图书馆关联数据孵化小组:数据集、属性值词汇及元数据元素集[EB/OL]. [2015-09-03]. [http://iskocn.org/lld/LLD\\_Datasets\\_Report\\_zh.html](http://iskocn.org/lld/LLD_Datasets_Report_zh.html).
- 12 223 Best Practices URI[EB/OL]. [2015-09-03]. [http://www.w3.org/2011/gld/wiki/223\\_Best\\_Practices\\_URI\\_Construction](http://www.w3.org/2011/gld/wiki/223_Best_Practices_URI_Construction).
- 13 INSPIRE Directive (2007/2/EC)[EB/OL]. [2015-09-03]. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF>.
- 14 Phil Archer, et al. Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC[EB/OL]. [2015-09-03]. <http://philarcher.org/diary/2013/uripersistence/>.
- 15 Stuart Williams. URI Patterns[EB/OL]. [2015-09-03]. <https://github.com/UKGovLD/URI-patterns-core/blob/master/URI%20Patterns.md#reference.URISetsV1>.
- 16 Australian Government Linked Data Working Group. URI Guidelines for publishing linked datasets on data.gov.au.v0.1[EB/OL]. [2015-09-03]. <https://github.com/AGLDWG/TR/wiki/URI-Guidelines-for-publishing-linked-datasets-on-data.gov.au-v0.1>.
- 17 British Library URI patterns[EB/OL]. [2015-09-06]. [http://www.bl.uk/bibliographic/pdfs/british\\_library\\_uri\\_patterns.pdf](http://www.bl.uk/bibliographic/pdfs/british_library_uri_patterns.pdf).
- 18 编目精灵.LC 关联数据服务:数据集现状(2015)[EB/OL]. [2015-09-06]. <http://catwizard.net/posts/20150825140838.html>.
- 19 French National Library, Semantic Web and Data Model[EB/OL]. [2015-09-06]. <http://data.bnf.fr/en/semanticweb>.
- 20 小河尘.URI 设计模式[EB/OL]. [2015-09-06]. <http://1drv.ms/1XsAHEw>.
- 21 ANDS Guides. Persistent Identifiers[EB/OL]. [2015-09-06]. <http://ands.org.au/guides/persistent-identifiers-working.html>.

(许磊 馆员 上海图书馆,夏翠娟 高级工程师 上海图书馆,刘炜 研究员 上海图书馆副馆长,张磊 高级工程师 上海图书馆)

收稿日期:2016-08-26