# 基于本体的数字图书馆信息资源构建

□张敏勤

**摘要** 根据数字图书馆中不同的信息资源，从宏观上建立基于本体的数字图书馆信息资源构建的3个层次结构，即基于本体的文献信息资源构建、Web信息资源构建及知识管理中知识库的构建等，并从微观上为每一层次提出可操作的方法体系。

**关键词** 数字图书馆 信息资源构建 本体论

数字图书馆信息资源包括许多层次，如文本文献信息资源、多媒体信息资源及知识管理仓库等，因此，基于本体的数字图书馆信息资源构建是一个多层次的体系，包括以下几个层次的信息构建。

## 1 基于本体的文本文献信息资源构建

利用本体分析信息资源间的内在关系，按学科领域进行分类和索引，并丰富其语义，建立本体驱动的分布式学科信息门户，以支持对学科信息的有效搜索和检索。这是基于本体的信息构建的主要目的。文本文献资源是传统图书馆馆藏的主体，随着图书馆数字化进程的深入，对揭示大量文本文献内在关系的需求更为紧迫，对于文本文献来说，构建本体论的实质就是建立文献之间的等级结构，并定义文献之间的关系。文献的创造过程始于一种思维活动，即“构思”。“构思”经过抽象，形成一定形式的“表述”。将“表述”实体化，就成为“版本”。“版本”通过“载体”而存在；如果载体是数字化的，则称为“数字化载体”。“载体”即为文本文献资源中的本体“实例”[1]。

文献等级各层次之间是一种继承关系。“实例”是“载体”或“版本”中的一种，因此，给“实例”一个统一题名，这一统一题名可以直接检索到各文献等级，而无需利用文献等级结构导航。众多的“实例”形成一个以全局本体为语义的共享词表，所有信息源都与某一全局本体关联。在文献著录逻辑中，文献等级中的每一个概念均“属于”其上位类，而“延伸”至其下位类。继承的值与“延伸”相关上位类的值“相同”。这种等级关系在MARC上有较好的体现，所以利用MARC数据进行文本文献信息构建本体具有较强的可行性和实用性。将MARC字段和值转化为本体标识要借助4个控制文档，这组文档著录了MARC格式及其本体标识。主控文档将选定的MARC字段和编码扫描为一个或多个本体概念。根据不同的MARC值，确定将单一字段扫描为多个本体概念。第二个控制文档对多个MARC字段扫描为同一本体概念的情况建立优先权。第三个控制文档依据MARC记录的类型和书目层次标识MARC字段内编码的位置。第四个控制文档记录每一个编码的信息，包括编码值表的长度和位置[2]。构建本体论的最终目的是应用，生成可供利用的文本文献本体知识库是基于本体的文本文献构建的最后一步。将MARC数据转为本体概念标识的文本，依据本体论所规定的文献等级结构及定义的相关属性可以进行如下推理：

● 合并相同的构思。有相同或相似题名、作者、主题词的构思被视为相同的构思。给相同的构思一个统一题名，标识、合并在一起。

● 合并具有相似文风或相关作者值匹配的构思，由版本层的某种衍生关系可以推理出相似构思并在较好识别的情况下，标识和合并具有相似构思的表述。

● 合并出版者和出版日期相匹配并具有等级关系的版本项，如版本项之间存在隶属关系或存在载体层某种衍生关系，可以标识和合并具有相同或相似表述的版本项。

● 用统一的标准对主题概念的多项进行合并。统一标准算法使用基数限定，这一基数限定是本体

论定义的一部分，它支持推理演算。如果某项的类型主题属于另一项，且允许多重值，那么该项被合并。

●依据 MARC 记录描述的衍生关系建立与先前文献的关系。如文献各种版本间关系、衍生文献与原著间关系等生成原作相关知识的本体论。

最后将推理的结构进行归纳，将具有相同关系的文献合并，从而生成不同的知识库[3]。

## 2 基于本体的 Web 信息资源构建

数字图书馆的信息资源不仅包括文本文献，还包括数据库、Web 信息资源等多种多媒体信息，网上信息资源虽然从局部看可能是良构的，但从总体上看却是无序和劣构的，因为它们散布在 Internet 的数百万台服务器中，缺乏统一的分类和编码。而且资源及其使用环境具有异质性，这种异质性会引起互操作问题。互操作问题会出现在四个不同层面上：系统层、文法层、结构层及语义层。系统层有硬件和操作系统的不兼容问题，文法层有不同语言和数据表达问题，结构层有不同数据模式问题，语义层有信息交换中术语的意义差别问题。因此 Web 信息所固有的异构性、多分布性、增长性和变化性决定了在信息资源构建中结构方法不再适合，语义方法是当前多媒体信息资源构建研究的重点。正如目前网络协议的分层原理一般，基于本体的语义网表示可以满足不同用户群的需要，提供简单的分类方法和关系，附加层的表达性、功能性和复杂性可根据不同的用户需求增加，从而实现可扩展性和语言的表达性之间的平衡[4]。基于本体的语义网可实现以下层次的功能：

(1)XML 提供结构化文档的基本格式，具有灵活性和可扩展性优势，突破了 MARC 的局限，可以描述各种类型的信息资源。

(2)RDF 模型和语法规范提供引用的概念，基本断言模型提供概念属性，“实例”为概念或属性中的一种，只要给出实例的语义表示，即可生成类实体——关系模型。

(3)本体证明语言 RDF/RDF(S)是一种基于 XML 的标准化、互操作语言，其数据模型与语义网络形式等值。它允许发送一个代理到另一属性，可实现不同结构信息资源间的统一检索。

(4)本体模式语言 SHOE 和基于 SHOE 的本体标识语言 OML 提供包含分类目录和关系规则的分布式本体，使得 Agents 能够收集有意义的 Web 页面的文档信息。

(5)本体转换语言 XOL/OIL/OWL/OWL－S 允许推理、演化规则的表示，使用推理函数可将一种限定的声明语言变为一种完全图灵式的逻辑语言，而演化规则语言允许具有某种算法的机器将文档从一种 RDF 模式转换成另一种 RDF 模式，增强了本体语言的表达和检索功能。

(6)本体检索语言 RQL/DQL 等支持不同形式的查询引擎，同时具有相关的推理能力，可以通过规范语言标准化某些查询引擎和定义查询引擎[5]。

传统的 Web 信息资源以“知网”为分类体系，知网是一个揭示概念与概念之间以及概念属性之间的关系为基本内容的常识知识库，概念及属性之间的各种关系包括：上下位关系、同义关系、反义关系、对义关系、属性—宿主关系、部分—整体关系、材料—成品关系、事件—角色关系等等[6]。基于本体的 Web 信息资源构建中等级关系的建立，可直接将“知网”转化为本体，知网中的等级关系即本体论中的等级关系。同时由于本体语言中的推理功能，本体论可以推断不完整的知识。因为继承关系是本体论的核心，而目前的网络标识语言 XML 本身并不支持这种关系，因此需要借助其他方法在 DTD 中生成。使用 XML 的参数实体可以实现这一目标，参数实体定义了可以用于 DTD 的替换字符串，每当参数实体被参照时，这一参照则使用替换字符串来代替。DTDMAKER 是一个在 XML 文献的 DTD 中构建本体论的有效工具，它将本体论的概念扫入 DTD 的元素类型中，即对每一个概念元素类型都做定义，这些元素类型的内容模型由表达概念属性的元素构成[7]。等级关系的确定即形成了基于本体的 Web 信息资源分类体系，其知识库的生成只要通过对本体的维护即可完成，知识库的生成要与分类体系相结合，首先根据某种方法建立一个类别的本体论体系，然后根据专业知识对该类的本体概念进行修改，随着网络信息资源的不断增加，搜索引擎得到更多的查询内容，建立一个不断更新的领域知识库[8]。

## 3 基于本体的知识管理知识库构建

数字图书馆的目标是在对各种文献文本信息进行组织和管理的基础上，最终实现对知识的管理。在知识管理的全过程中构建本体论，可以实现对知识本身的揭示，实现数字图书馆中信息资源最高层次的构建。

知识的开放性集成主要包括三部分功能：知识转换、知识抽取和知识标识。通过这三部分的处理，将各种信息源中的知识按本体规定的形式表达，形成结构化的知识存储到知识库。因此，在知识管理的过程中构建本体论主要由三项工作构成：(1)获取知识，根据“知网”建立等级结构；(2)按本体规则对知识进行描述、存储，以形成知识库；(3)在推理基础上提供知识的智能检索，以实现知识重用[9]。

知识的获取是指在知识信息上加上元数据，将结构化、半结构化和非结构化信息转化为结构化信息的过程，从知识管理的角度，信息来源于以下知识源：人员或团体的背景信息；人力资本的隐性知识；各种同构的或异质的数据库；各种文献文档资料；从Web上挖掘出的知识等。

对获取的知识进行本体描述，可通过以下步骤完成：对于结构化的数据，在HTML中对被描述的信息客体添加本体论的onto语句，在结构化的数据库或标准的Web浏览器中，onto语句的添加并不影响HTML文件的视觉效果，这样做可以使得主题事物知识的查找可视化。半结构化和非结构化的数据，可利用XML中的DTD自行定义所需的标记语言及XML文件的结构。利用OML对分布在网络环境下需要管理的知识资源添加本体论的标识语句，在标识中可直接使用(或再利用)语句体中的文本知识，避免知识标注者重复表示相同的信息。将上述转化以后的知识，包括结构化的元知识和半结构化和非结构化的信息体分别存放于数据库和文档库中[10]。

知识的重用即知识的应用过程，在知识管理中就是实现人与知识的连接过程。主要包括四种连接：(1)连接人到知识，指通过可视化的查询与检索工具，从不同的信息源中找到相关内容，以pull的方式获取知识。在网络环境下可以使用基于本体论的代理服务Ontobroker，它由三个部分组成：网络爬虫、推理引擎及查询界面。首先，Ontocrawler通过标识的网页进行查找并收集标注的知识片断。然后，将知识片断以Ontocrawler所使用的表述语言进行规范。推理引擎收到用户的提问后，利用两个信息源来推导答案，即主题事物的本体论和Ontocrawler中的事实。推理引擎的基本推理机制类似于知识库中的智能推导系统。形成知识的智能检索系统。(2)连接人到人，知识管理者，如图书馆参考咨询馆员可通过KM系统找到相关领域的专家咨询或讨论。(3)连接知识到人，即知识管理中的个性化推送服务，根据用户的个人偏好或需要，主动推送相关知识。(4)连接知识到知识，

由于在本体标注和查询的过程中使用了本体论标识语句，本体论元数据表示的是概念的等级关系，这种等级关系至少容易满足用户的两种需求：其一是浏览某一类的周围类目，以便寻找最合适的形成某一提问的类，其二是浏览全部等级，以便快捷地完成由一个等级向另一个等级的导航。因此实现了知识之间的超链接[11]。

## 参考文献


1 数字图书馆信息资源本体论的构建. 2006－01－19.[2006－7－18]. http://www.duozhao.com/lunwen/j17/lunwen_59762.html

2 同1

3 刘柏嵩. 面向语义网的本体表示. 中国图书馆学报，2004(2)：47－50

4 同3

5 李景. 主要本体表示语言的比较研究. 现代图书情报技术，2005，(1)：1－4，8

6 凌云，魏贵义，刘军. 基于Ontology的Web文本分类法. 情报学报，2005(2)：202－207

7 同1

8 丁晟春，岑咏华，顾德访. 基于Ontology的语义检索研究. 情报学报，2005(6)：702－707

9 刘柏嵩. 基于本体的知识管理关键技术研究. 情报学报，2005(1)：75－81

10 董慧，杜文华. 基于本体和多代理的数字图书馆信息检索模型. 中国图书馆学报，2004(2)：63－65

11 同9


作者单位：安徽工业大学图书馆，马鞍山，243002
南京大学信息管理系，南京，210093

收稿日期：2006年7月20日

(下转第92页)

## 参考文献

1 罗冰眉. 网络环境下个人数据与其隐私权的保护. 现代情报，2003(9):55

2 周德堂. 个人信息及其保护. 法制研究，2005(8):192

3 刁胜先. 论网络隐私权之隐私范围. 西南民族大学学报(人文社科版)，2004(2):2

4 齐爱民. 中华人民共和国个人信息保护法示范法草案学者建议稿. 国家社会科学基金项目(04XFX006). 2005 年 2 月 27 日:2

5 刘绿茵. 现代图书馆文献信息服务的十大趋势. 中国信息导报，2005(12):23－25

6 赵英. 数字图书馆建设中的隐私权保护. 四川图书馆学报，2005(3):42

7 侯巍. 联合国对个人资料的国际保护. 广西大学学报(哲学社会科学版)，2005(4):62

8 张新航，张廷川. 我国图书馆立法的必要性及其对策. 贵州民族学院学报(哲学社会科学版). 2005(3):122

9 吴育珊. 网络个人信息知情权保护机制探析. 经济与法·南方经济，2005(5):29

10 张军，熊枫. 网络隐私保护技术综述. 广东省自然科学基金资助项目(020199)，2004 年 4 月 17 日:11


作者单位:华南师范大学经济与管理学院，广州，510006

收稿日期:2006 年 7 月 21 日

# Effects and Solutions of the Law of Personal Information Protection to Information Service of library

Liu Qing　Huang Yuanyuan

**Abstract**: This article analyzes the change of information service of Chinese libraries caused by the law of personal information protection, which is nearly confirmed. Based on the principles of the law, the authors analyze its effect to the whole process of information service and finally discuss the solution the libraries should take to face the challenge.

**Keywords**: Law of Personal Information; Library; Information Service; Effect; Solution

(上接第 45 页)

# The Information Resources Construction of Digital Library Based on the Ontology

Zhang Minqin

**Abstract**: This article puts forward a kind of lately solution of Digital Library information resources construction—ontology. If introduces the concept of ontology and Ontology Digital Library and describes this concept from the aspect of metadata. According to the difference kinds of Digital Library information resources, the paper sets up three layer frameworks on ODL, that is documental information construction on ODL, Web information construction on ODL and knowledge database on KM construction on ODL. Finally it gives out a set of workable methodology.

**Keywords**: Digital Library; Information Resources Construction; Ontology