数字图书馆元数据基础

SHUZI TUSHUGUAN YUANSHUJU JICHU

萨蕾／著

中央编译出版社
Central Compilation & Translation Press

数字图书馆元数据基础

SHUZI TUSHUGUAN YUANSHUJU JICHU

萨蕾／著

中央编译出版社
Central Compilation & Translation Press

图书在版编目（CIP）数据

数字图书馆元数据基础／萨蕾 著

—北京：中央编译出版社，2015.8

ISBN 978-7-5117-2746-6

Ⅰ.①数… Ⅱ.①萨… Ⅲ.①数字图书馆－研究

Ⅳ.①G250.76

中国版本图书馆 CIP 数据核字(2015)第 184780 号

数字图书馆元数据基础

出 版 人：刘明清

出版统筹：董 巍

责任编辑：王丽芳

责任印制：尹 珺

出版发行：中央编译出版社

地 址：北京西城区车公庄大街乙 5 号鸿儒大厦 B 座（100044）

电 话：（010）52612345（总编室） （010）52612349（编辑室）

（010）52612316（发行部） （010）52612317（网络销售）

（010）52612346（馆配部） （010）55626985（读者服务部）

传 真：（010）66515838

经 销：全国新华书店

印 刷：北京京华虎彩印刷有限公司

开 本：787 毫米×1092 毫米 1/16

字 数：186 千字

印 张：13.25

版 次：2015 年 8 月第 1 版第 1 次印刷

定 价：68.00 元

网 址：www.cctphome.com **邮 箱**：cctp@cctphome.com

新浪微博：@中央编译出版社 **微 信**：中央编译出版社（ID：cctphome）

淘宝店铺：中央编译出版社直销店（http://shop108367160.taobao.com）

（010）52612349

导　言

随着数字图书馆的发展，数字资源逐渐成为公共图书馆馆藏不可或缺的一部分。元数据一直是图书馆实现文献有序化的主要工具，同样，在数字图书馆的建设中，元数据也将起到重要的作用。

与图书馆传统编目工作相比，数字资源元数据的建设有其自身的特点。

一、数字资源具有海量的特点，因此，对数字资源进行编目不能如传统编目工作那样完全依赖图书馆员，而需要结合自动编目方法。自动编目可以极大地节省人力，提高元数据制作的速度，但是，数字资源、尤其是网络信息资源的自由性、不规范性、非结构化特点会影响元数据的质量，因此，数字资源元数据的建设应加强对信息自动处理技术的研究，充分利用受控词表、本体等知识组织工具提高对文本进行自动处理的准确性。

二、公共图书馆的数字资源建设主要有两个途径。不同的建设途径也使元数据具有不同的特点。

1. 对馆藏实体文献进行数字化加工形成数字文献（在本书中称为自建数字馆藏）。

自建数字馆藏由于来源于实体文献，因此，其元数据的建设应充分利用图书馆已建的书目数据，一方面利用了已有的知识组织产品，节省了人力物力，另一方面，在制作过程中可以同时建立数字化文献和印本文献间的关联关系。

多年来，图书馆的编目工作一直以手头文献为著录对象，而数字文

献由于有了数字文本，就有了对信息资源的内容乃至知识点进行挖掘与组织的可能。因此，数字资源元数据除了可以以文献为著录对象外，还应进一步以作品为著录对象，摆脱文献由物理形态带来的限制，围绕作品整合资源，最终实现整合知识，建立多层次的关联关系，为用户提供准确高效的知识导航。

2. 按照本馆的职能与服务需求对网络信息资源进行采集，纳入本馆的馆藏。

网络信息资源与图书馆传统馆藏有着很大的区别，因此，建设网络信息资源元数据应首先研究制订适用于网络信息资源的元数据标准体系，标准的制订既要关注网络信息资源的特点，又要与通用的元数据标准保持相对的一致，以保证标准具有开放性，能通过互操作实现信息资源的共建共享。

三、数字资源具有多种来源、多种媒体类型，因此，对数字资源元数据的利用中很重要的一个环节是整合不同来源、不同类型的元数据，揭示其中的关联关系，使用户可以一站式获取所需资源。

图书馆的数字资源元数据整合应根据本馆馆藏情况、系统建设的特点来考虑元数据整合的方式。一种方式是基于 OPAC，以实体文献的书目数据为核心，整合相关数字资源元数据，这种方式适用于数字资源建设初期本馆馆藏仍然以书目数据为主的情况。另一种方式是构建元数据仓储，整合各种类型、各种来源的元数据，实现一体化的元数据服务，这种方式适用于建设了大量不同来源、不同格式的数字资源元数据的情况。

四、数字资源的海量性、我国不同地区图书馆建设水平的差异性使得很多图书馆没有能力独自建设数字资源，而不论是数字图书馆的整体发展需求还是由社会发展带来的用户需求都需要各类型公共图书馆尽快加入到数字资源的建设中来，因此，公共图书馆的数字资源元数据建设应以共建共享为主要模式。

数字资源元数据的建设对于图书馆来说，还是一个需要不断研究、

不断实践、不断发展的新兴领域。如何基于数字资源元数据的特点，更好地实现元数据揭示数字资源内容、提供多维知识服务的功能是本书的研究重点。

本书内容共分五部分。第一部分包括第一章，对元数据的基础理论进行了研究。介绍了元数据的概念、类型、功能等，并对 MARC 元数据、DC 元数据这两种图书馆界最通用的元数据格式进行了研究，并对元数据在语义网环境下的发展进行了论述。第二部分包括第二章至第五章，研究了自建数字馆藏元数据的建设。将著录对象分为三个层次——文献、作品、知识对自建数字馆藏元数据进行了深入的研究，并按照这三个层次以古方志元数据作为实例研究自建数字馆藏元数据的建设。第三部分包括第六章至第八章，研究了网络信息资源元数据的建设。基于网络信息资源的特点研究元数据建设原则及实现途径，并通过研究网络政府信息元数据标准体系、受控词表，对图书馆网络政府信息资源的建设进行了全面的探讨。第四部分包括第九章至第十一章，在前面各章研究的基础上，探讨基于元数据实现数字资源整合的方式。第五部分包括第十二章，提出应以共建共享为公共图书馆数字资源元数据建设的模式，并系统论述了合作体系的组织架构和平台建设。

由于研究能力与研究时间有限，因此对一些问题的研究不够全面与深入。对于错误与疏漏之处，敬请各位专家学者及图书馆同仁不吝赐教。

萨蕾

2015 年 4 月

目　录

第一章　元数据标准研究

第一节　元数据概述

一、元数据的定义

元数据最通俗的定义是“关于数据的数据”，但是，这一定义太过宽泛、模糊，因此，多年来，对元数据的研究往往是从元数据的定义入手，从而产生了多种更明确更细化的定义。包括：

data that defines and describes other data. [ISO/IEC 11179 - 3:2003 (E)]

元数据是关于数据的数据。此术语指任何用于帮助网络电子资源的识别、描述和定位的数据。(IFLA)

用于提供某种资源的有关信息的结构数据（Structured data），或者说是描述其他数据的数据（Data about other data）。[①]

元数据是用于描述数据内容（what）、覆盖范围（where when）、质量、管理方式、所有者（who）、提供方式（how）的数据，是数据与数

① 王松林：《论网络信息资源的元数据编目》，载《图书馆学刊》2004年第2期，第9—11页。

据用户之间的桥梁。①

元数据是与对象相关的数据，此数据使其潜在的用户不必预先具备对这些对象的存在或特征的完整认识。它支持各种操作。用户可能是程序，也可能是人。②

元数据是对信息包（Information package）的编码描述，其目的在于提供一个中间级别的描述，使得人们据此就可以做出选择，确定孰为其想要浏览或检索的信息包，而无需检索大量不相关的全文文本。③

关于信息资源或数据的一种结构化的数据。④

是面向某种特定应用的用于描述资源属性的机器可理解的信息。通过规范语法结构和语义结构，使得机器能够无二义性地表现和获取信息。⑤

对上述定义进行分析，可以看到，这些定义对元数据的特征进行了概括，包括以下几点：

（一）元数据的描述对象

元数据的描述对象包括数据、信息或知识。数据是对客观事物、事件的记录、描述，是可由人工或自动化手段加以处理的数字、文字、图形、图像、声音等符号的集合。信息是客观世界中各种事物的状态和特征的反映，是与问题相关的数据，可以以文本、图形、图像、音频、视频等形式记录下来，能通过媒介进行传输。知识是人们从实践经验中总

---

① 刘炜：关于元数据的十万个为什么[EB/OL].[2015-04-20]. http://www.libnet.sh.cn/sztsg/fulltext/abc/metaFAQ.pdf.

② Dempsey, Lorcan and Herry, Rachel. Metadata: a current view of practice and issues. The Journal of Documentation, vol. 54, no. 2 (March 1998): 149, 157.

③ Taylor, Arlene G. The organization of information. Eaglewood, Colorado: Libraries Unlimited, Inc., 1999, 246.

④ 肖珑、申晓娟：《国家图书馆元数据应用总则规范汇编》，北京：国家图书馆出版社2011年版，第3页。

⑤ 梁蕙玮：《公共图书馆自建资源整合研究与实例分析》，北京：国家图书馆出版社2014年版，第37页。

结出来且被新的实践所证实的规律及经验的总结，是可以用于推理的规则。[①]

笔者认为，不对描述对象进行限制更能体现元数据的作用，也更符合元数据的本质特征。基于此认识，元数据所描述的对象既包括实体文献，也包括数字资源；既包括单个的独立资源，也包括集成性资源（如：期刊、网站等），以及多媒体资源等；既包括物理实体，也包括虚拟实体；既包括知识组织工具（如：词表）、服务、信息系统等，也包括元数据本身。而从数据的类型上看，包括了书目型数据、文献型数据、数字型数据、数值型数据等。

（二）元数据涵盖的范围

随着因特网的发展，元数据的概念出现在图书情报领域，一般认为，元数据这一词汇最早出现在1988年美国国家航空航天局（NASA）的《目录交换格式》（Digital Information Formats，DIF）手册中。在这一词汇出现之前，图书馆已经进行了多年的编目工作，制作了大量的书目数据、索引数据等。最为通用的IFLA的定义也将元数据的描述对象局限于网络电子资源。但是，也有学者认为书目数据是元数据，TEI标题也是，或其他形式的描述。[②] 因此，笔者认为，尽管描述对象发生了很大的变化，使得描述网络资源的元数据与传统的书目数据在元数据标准上有着一定的差异，但是，从元数据的描述对象来看，传统的书目数据、目次数据等也应是元数据的一部分，描述网络资源的元数据与书目数据仍然是同一性质，同时，将元数据局限于对网络电子资源的描述不利于元数据的发展。因此，对元数据的定义应是较为宽泛的，从元数据所起的作用出发进行定义更为科学合理。

---

① 郑彦宁、化柏林：《数据、信息、知识与情报转化关系的探讨》，载《情报理论与实践》2011年第7期，第1—4页。

② 刘嘉：《元数据导论》，北京：华艺出版社2002年版，第43页。

（三）元数据的属性

元数据最基本的属性包括两点：

1. 元数据具有明确的语义和结构。

元数据具有明确的语义和结构，这一性质使得元数据互操作具有了可行性。元数据的互操作性体现在对异构系统间互操作能力的支持。基于互操作实现的信息资源的共享与融合可以使元数据发挥更大的使用价值，因此，互操作性在元数据使用与发展的过程中具有重要意义。

然而，不同的元数据都具有一定的个性化特点，因此，在互操作时，应以损失最小化为原则。

2. 元数据是机器可读、可理解的数据，即：元数据的使用者可能是人，也可能是机器。在数字时代，这一点是元数据存在的基础，也是更好发挥元数据功能的重要保障。这一点也决定了元数据在生成、组织、管理、利用、保存等过程中保证与机器的语义互通是至关重要的。

（四）元数据的特点

元数据最显著的特点为描述性及模块化两点。

1. 描述性

描述是元数据最重要最基本的功能，不论是描述元数据还是管理元数据，都是在对描述对象的某一方面的属性进行描述。

描述性体现在元数据的形成过程中，其重要性在于只有加强描述性才能实现元数据管理、使用的功能，因此，在制作元数据时，要遵循客观性、准确性、完整性等描述原则。

2. 模块化

模块化（Modularity），指按照所描述的信息系统内容，将元数据划分为针对不同层次、功能或应用的逻辑模块，每个元数据格式只是一个这样的模块，分别对信息系统的不同内容进行描述，分别满足不同的逻

辑功能和应用需要。①

模块化主要以两种方式体现：一、不同的元数据类型形成了不同的模块。如：MARC 元数据标准家族包括了书目数据、馆藏数据、规范数据。二、不同的功能形成了不同的模块。如：国家图书馆元数据规范中就要求，元数据应包括描述信息、技术信息、管理信息。

模块化的优点在于：模块化使得元数据标准的建设具有动态性，一个模块可以独立使用，也可以实现按照具体需求与其他模块的多次重组；对于实际应用，往往需要多个模块的组合才能满足需求，模块化与资源描述与使用的客观需要相吻合；模块化有利于对元数据标准的复用，在复用时，可从其他元数据中选择部分模块直接复用，或按照自身的需求进行一定的扩展。

（五）元数据的结构

元数据包括内容结构、语法结构、语义结构。内容结构定义了元数据的构成要素，如：描述性元素、管理性元素、元素选取使用规则等。语法结构定义了元数据的格式结构及其描述方式，包括：元数据的结构、元素结构、元素复用方式、与描述对象的捆绑方式等。语义结构：语义结构定义了元素定义、元素内容编码规则定义等。语义通过属性元素表达，结构是语义的抽象载体，可以提供人类和机器的双重理解；语法（句法）是置标方案，用以传达语义和结构。

二、元数据的类型

研究者对元数据的类型也有多种认识。主要有以下几种：

（一）杨超认为数字图书馆领域的元数据可以分为三类：②

描述性元数据（descriptive metadata）：描述对象知识内容的信息，

---

① 张晓林：《数字图书馆理论、方法与技术》，北京：北京图书馆出版社 2007 年版，第 88 页。

② 杨超：《数字图书馆描述性元数据仓储模型研究》，上海：上海交通大学硕士学位论文，2009 年，第 12 页。

例如 MARC 编目记录，检索工具或者类似的模式、框架。

管理性元数据（administrative metadata）：允许仓储管理对象的必须的信息：包括扫描信息，储存格式等（通常这类元数据也被称作技术元数据），版权和许可信息，数字对象长期保存所需要的各种信息（保存元数据）。

结构性元数据（structural metadata）：将对象互相连接以形成逻辑单元的信息，例如一本书中每页内容的镜像互相关联从而形成这本书自身。

（二）Anne J. Gilliland-Swetland 将元数据分为五种类型：[①]

管理型：是在管理信息资源中利用的元数据。

描述型：是用来描述或者识别信息资源的元数据。

保存型：是与信息资源的保存管理相关的信息。

技术型：是与系统如何行使职责或元数据如何发挥作用相关的元数据。

使用型：是与信息资源利用的等级和类型相关的元数据。

（三）根据元数据在组织信息资源的功能上划分，元数据可分为：[②]

1. 知识描述性元数据（Intellectual Metadata），用来描述、发现和鉴别数字化信息对象，如 MARC、DC，它主要描述信息资源的主题、内容特征，体现在所形成的记录上，每条记录都是对数据值和内容的表达。

2. 结构型元数据（Structural Metadata），描述数字化信息资源的内部结构，如层次、类属、目录、章节、段落的特征。

3. 存取控制型元数据（Access Control Metadata），用来描述数字化信息资源能够被利用的基本条件和知识产权特征，以及这些资源的期限

---

① A. J. Gilliland-Swetland, Setting the Stage: Defining Metadata, Murtha Baca, Introduction to-Metadata: Pathways toDigital Information, Los Angeles: Getty Information Institute, 1998, 6 - 8.

② GB/T21063.3 - 2007，政务信息资源目录体系第 4 部分：政务信息资源分类［S］。

和使用权限。

4. 评价型元数据（Critical Metadata），描述和管理数据在信息评价体系中的位置。

（四）从元数据的层次体系角度划分元数据类型：

赵军认为：[①] 广义元数据的层次体系由第一层至第六层依次为：信息内容格式元数据、内容对象元数据（狭义元数据）、资源集合元数据、管理与服务机制元数据、过程与系统元数据、宏元数据。

肖珑、赵亮认为：[②] 根据数字资源从产生到服务的生命周期、元数据描述和管理内容的不同以及元数据作用的不同，分为多种类型：内容元数据：描述数字对象内容及结构的元数据。专门元数据：描述单一数字对象（如学位论文、古籍、网络资源、期刊论文等）的内容、属性及外在特征的元数据。资源集合元数据：按照学科、主题、资源类型、用户范围、生成过程、使用管理范围等形成的信息资源集合（如数据库、知识组织系统等）的描述。管理元数据：数字对象的加工、存档、结构、技术处理、存取控制、版权管理以及相关系统等方面信息的描述。服务元数据：数字资源服务的揭示与表现、服务过程、服务系统等方面的相关信息的描述。元元数据：对元数据的标记语言、格式语言、标识符、扩展机制、转换机制等信息的描述。上述体系中，信息内容元数据、专门数字对象元数据和资源集合元数据更多的是发挥对资源本身的内容、属性、外在特征的描述作用，称为描述元数据，另外几种则比较多地发挥了对资源和元数据的管理作用，称为广义上的“管理元数据”。

仔细考察上述的各种观点，可以看到，元数据的类型以描述型元数据与管理型元数据为主。对于描述型元数据基本上都有共识，而对于管

---

① 赵军：《数据资源描述与组织的元数据方法》，天津：天津大学硕士学位论文，2005年，第9页。

② 肖珑、赵亮：《中文元数据概论与实例》，北京：北京图书馆出版社2007年版，第11页。

理型元数据的认识，则有狭义与广义之分，狭义的理解是将管理功能与保存功能、技术特征和使用功能区分开，而广义的理解是将服务型元数据、技术型元数据、保存型元数据、存取控制型元数据等都作为管理元数据的一种。细化元数据的分类有利于按照不同的功能需求设置元数据的结构，而将管理元数据理解为广义概念，有利于对元数据进行管理和交换。

三、元数据的功能

对于元数据功能的认识一直随着图书馆发展史而不断发展着。

在西方，16 世纪以前的目录是作为财产清单来使用的，即目录只承担对文献的管理功能。到 17 世纪，文献的增加使得馆藏的文献数量增加，有了对文献进行查找的需要，于是，目录记录的内容逐渐复杂，查找功能开始出现。

1841 年，潘尼兹（Antonio Panizzi）主持制定了著名的《九十一条》，提出目录首先必须反映某一部特定的图书，但它并非把图书作为一个孤立的个体，而是把它作为具有特定著者的、某一著作的一个版本。这一理论除了阐述目录描述功能，还明确地提出了目录应具有对某一特定版本的查找功能，查找功能第一次在正式条例中得到体现。

1876 年，卡特（Charles Ammi Cutter）编制了《字典式目录规则》，概括了目录的功能：1. 使读者通过著者、题名、主题找到一本特定的图书；2. 揭示图书馆是否拥有一个特定著者、一个特定主题、某一特定文献类型的藏书；3. 帮助选择不同版本、不同特征的图书。[①] 这一阐述说明了目录功能从单纯的查检发展到同时具有检索与汇集功能。

1961 年，在巴黎的联合国教科文组织的会议大厦召开了国际编目原则会议，这是国际编目史上具有里程碑意义的国际会议。会议通过的

---

① Charle Ammi Cutter. Rules for a dictionary catalog. 4th ed. Washington: Govt. Printing Off., 1904:12.

《原则声明》对目录的功能做了这样的阐述：目录必须是一种有效的工具，用来确定：图书馆是否拥有某一特定的图书，如通过图书的题名、著者或其他合适的替代名称；图书馆藏有某一著者的哪些作品，或者藏有某一作品的哪些版本。至此，目录的两种功能在国际范围内被固定下来，成为编目界的统一思想，并延续了四十年。

到了20世纪末，巴黎原则声明的理念与信息资源及信息技术的发展有了一定的差距。图书馆界根据现实环境对目录功能进行了修订。

1997年，国际图联组织（IFLA）发布了《书目数据的功能需求》（FRBR），对目录功能做了这样的阐述：①

查找（find）符合用户检索要求的实体（即利用实体的属性或关系在一个文档或数据库中找到一个或一组实体）；

识别（identify）一个实体（即确认所描述的实体对应于所查找的实体，或者区分具有相似特征的两个或多个实体）；

选择（select）适合用户需要的一个实体（即选取一个在内容、物理形式等方面能满足用户要求的实体，或放弃一个不适合用户需求的实体）；

获取（acquire）或存取（obtain access）所描述的实体（即通过购买、借阅等方式获取一个实体，或者以电子方式通过联机连接远程计算机来检索一个实体）。

FRBR的阐述显示出网络环境对目录功能的影响：首先，著录对象涵盖了各种信息资源类型，包括文字、音乐、地图、视听、图形和立体资料；包含书目记录中描述的全部载体形态（纸质、胶片、磁带、光存储载体等）；包含各种格式（图书、单张出版物、唱片、双轴盒带、单轴盒带等）；而且反映所有记载信息的方式（模拟的、声学的、电学

① 书目记录的功能需求[EB/OL].[2015-04-20]. http://www.ifla.org/files/assets/cataloguing/frbr/frbr-zh.pdf.