

数字资源组织与揭示 ——元数据基础

元数据概述

什么是元数据？

- 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

- 关于信息资源或数据的一种结构化的数据。

——国家图书馆元数据应用总则

- 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。

DC元数据

概述

DC元数据（都柏林核心元数据， Dublin Core Metadata）

1995年3月，由OCLC与国家超级计算应用中心（NCSA）联合发起，因创始地在美国俄亥俄府都柏林而得名。目的是建立一套描述数字资源的标准。由DCMI负责维护

DC元数据

描述对象：

“资源（Resource）”。资源是任何可以标识的东西。可以是实体的，也可以是抽象的。常见的例子有电子文档，图像，服务（例如，“洛杉矶今天的天气预报”），还有其他资源的集合。并非所有的资源都是网上可检索的；例如，人，机构，还有图书馆里装订成册的书都可以被认为是资源。

（引自《关于元数据的十万个为什么》刘炜）

DC元数据

发展历史

- ◆ DC-1 (1995) : DC元数据产生。确定DC元数据的描述对象、13个元素, 确立各种基本原则。
- ◆ DC-2 (1996) : 讨论DC元数据应用的语法和结构问题, 提出Warwick框架作为元数据应用的一般“容器”和概念框架。
- ◆ DC-3 (1996) : 提出DC元数据在互联网图像资源描述方面的应用方案。增加了描述Description、权限Rights两个元素。

DC元数据

- ◆ DC-4 (1997) : 讨论DC元数据应用的扩展问题。提出三类“堪培拉限定词”，即对DC元数据元素可以从取值的language、Scheme和Type三个方面进行限定或扩展。提出了DC限定版。
- ◆ DC-5 (1997) : “荷兰终结”，指对DC的非限定版DCMES的最终确立。讨论了一对一原则。
- ◆ DC-6 (1998) : 形成了DCMI的基本组织形式和运行模式。

DC元数据

- ◆ DC-8（2000）：提出应用纲要形式解决元数据的领域应用问题。
- ◆ DC2003：提出抽象模型，对各类应用纲要进行规范。
- ◆ DC2007：提出了被称为“新加坡框架”的元数据应用规范。
- ◆ 至今，DC已发展成为涵盖一系列文档的标准规范体系，其中，除了以DC元数据术语集为主的语义规范外，还发展起包括词表、编码规范、模型等文档的一整套辅助规范。

DC元数据

特点

◆简单性原则

定义一个能得到最广泛应用、被全球所理解和接受的最小元素集。

◆内在本质原则(Intrinsicality)

集中于描述对象内在的属性。如：知识内容、物理形式。信息来源于描述对象，又不超出描述对象。

◆可扩展原则(Extensibility)

为描述独特的资源种类，因此必须有适当的弹性。可以基于DC核心元素集通过各种方式扩展为适应各领域资源描述需要的元数据方案。

DC元数据

- ◆语法独立原则(Syntax-Independence)

只规定元素的基本语义，元素可以以多种方式编码，避免语义的捆绑。

- ◆可选择性原则(Optionality)

所有的元素都是可选择的。

- ◆可重复原则(Repeatability)

所有元素均可重复。在此原则下，不区分作者的排名。

- ◆可修饰原则(Modifiability)

元素可独立使用，也可利用修饰词进行限定。限定不能扩大或改变元素的基本语义。

DC元数据

限定 (Refinements)

是对元数据语义的进一步限定和细化，也叫修饰词。包括元素修饰词、编码体系修饰词。

元素修饰词：对元素的语义进行修饰，提高元素的专指性和精确性。

编码体系修饰词：用来帮助解析某个术语值的上下文信息或解析规则。其形式包括受控词表、规范表或解析规则。起到对网络资源进行规范控制的作用。

DC元数据

术语：

摘要 (refines: 描述)、访问权限 (refines: 权限)、更新方法、更新周期、更新政策、交替名称 (refines: 名称)、适用对象、可获得日期 (refines: 日期)、文献引用 (refines: 标识符)、遵循 (refines: 关联)、其他责任者、覆盖范围、创建日期 (refines: 日期)、创建者、日期、接受日期 (refines: 日期)、版权日期 (refines: 日期)、递交日期 (refines: 日期)、描述、教育水平 (refines: 适用对象)、大小 (refines: 格式)、格式、格式转换为 (refines: 关联)、部分为 (refines: 关联)、版本关联 (refines: 关联)、标识符、指导方法、格式转换于 (refines: 关联)、部分于 (refines: 关联)、被参照 (refines: 关联)、被替代 (refines: 关联)、被需求 (refines: 关联)、发布日期 (refines: 日期)、版本继承 (refines: 关联)、语种、许可 (refines: 权限)、中介 (refines: 适用对象)、媒体 (refines: 格式)、修改日期 (refines: 日期)、保管历史、出版者、参照 (refines: 关联)、关联、替代 (refines: 关联)、需求 (refines: 关联)、权限、权利持有者、来源、空间 (refines: 覆盖范围)、主题、目录 (refines: 描述)、时间 (refines: 覆盖范围)、名称、类型、生效日期 (refines: 日期)

DC元数据

词表编码体系	句法编码体系
DCMIType	Box
DDC	ISO3166
IMT	ISO639-2
LCC	ISO639-3
LCSH	Period
MESH	Point
NLM	RFC1766
TGN	RFC3066
UDC	RFC4646
	RFC5646
	URI
	W3CDTF

DC元数据

DC元数据应用纲要（DCAP）

纲要是描述一个标准或一些规范是如何被运用以支持特定的应用、功能、行业需求或特定环境的文档。元数据纲要是一种元数据标准的应用形式，也可以看成是一种规范的元数据方案。

元数据纲要的作用：一、保证了元素基本集的稳定，二、实现了特定应用领域对元数据标准的需求，三、对DC元数据标准的应用进行了一定程度的规范。

DC元数据

2001年《DC图书馆应用纲要》（草案）发布

元素：18个

元素限定：26个

编码体系修饰词：21个

术语属性包括：

术语名称；术语URI；标签；定义来源；初始定义；DC-Lib定义；
初始注释；DC-Lib注释；术语类型；限定；被限定；编码体系应
用于；可应用的编码体系；约束；出现次数

DC元数据

元素	元素修饰词
题名	交替题名
创建者	
其他责任者	
出版者	
主题	
描述	摘要、目录
日期	创建日期、生效日期、可获得日期、发布日期、修改日期、版权日期、递交日期、接受日期、获取日期
类型	
格式	大小、媒体

DC元数据

元素	元素修饰词
标识符	文献引用
来源	
语种	
关联	版本继承、格式转换于、格式转换为、被替代、替代、部分于、部分为、需求、被参照、参照
覆盖范围	空间、时间
权限	
适用对象	
版本	
馆藏位置	

DC元数据

DC元数据抽象模型 (DCMI Abstract Model, DCAM)

在一套概念术语的基础上，提供一个抽象的数据模型。定义了DC元数据描述的各类实体对象及其相互之间的关系，描述DC元数据所使用的描述资源的信息结构。提供了一种独立于任何特定编码方式的信息模型，实现在不同元数据方案之间的共享和重用。

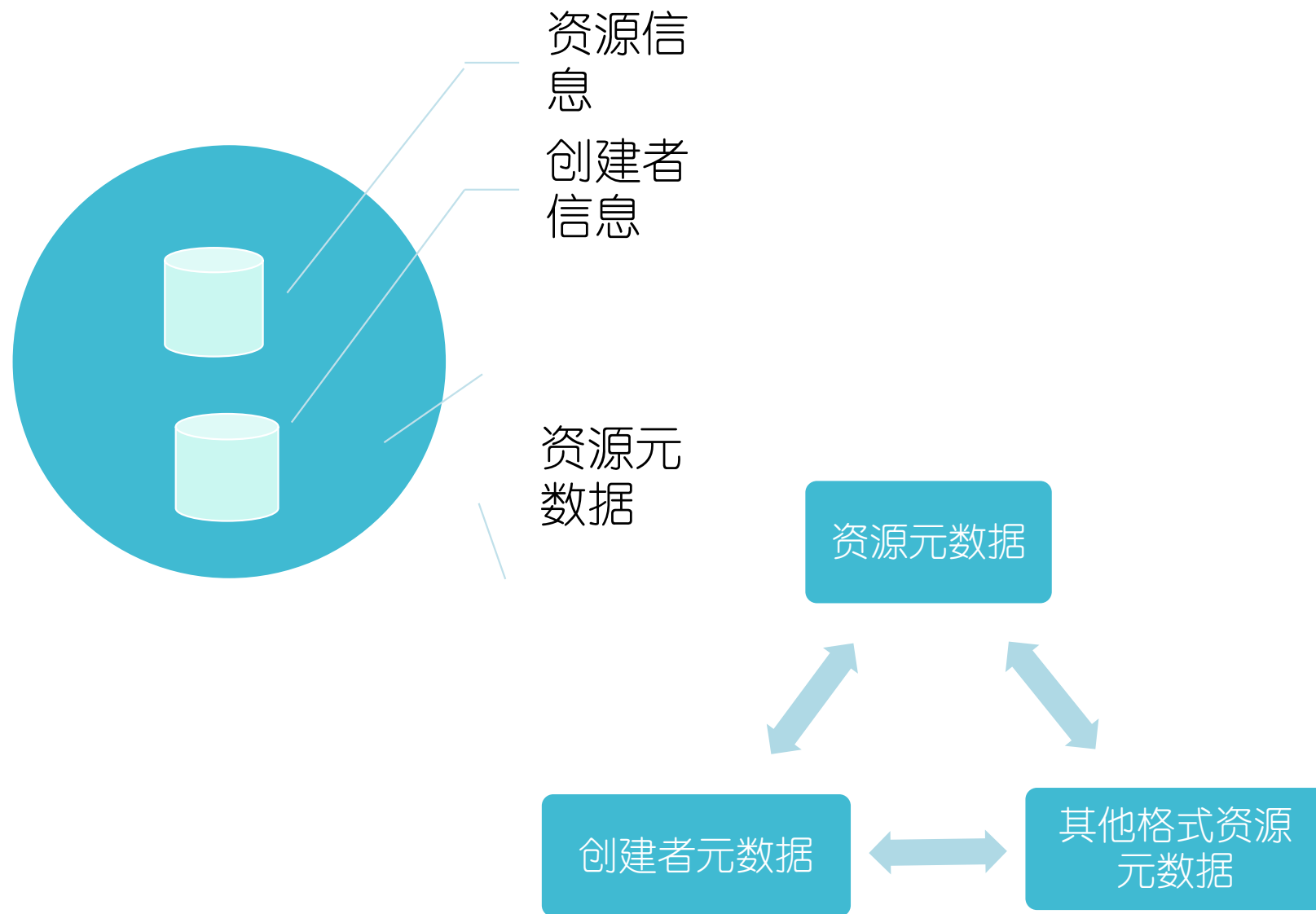
- 2005年，DC元数据抽象模型成为DCMI的推荐规范
- 2007年，发布修订版本。

DC元数据

原则1：一比一原则（One-to-one Principle）

一条描述包括了一个或多个陈述，而这些陈述与一个并且仅仅一个资源相关，即：一条描述只能描述一个并且仅仅一个资源。

DC元数据



DC元数据

原则2：向上兼容原则（Dumb-down principle）

词表模型中规定，如果一个“属性—值”对使用子属性及对应的值来描述一个资源，那么这个资源也可以用子属性关联的属性（父属性）及子属性的值来描述。

实例

请为一个资源库设计一套元数据标准

<http://www.tjwh.gov.cn/shwh/lywh/mrgj/gu-wei-jun/RWJS-gwj.HTM>

MARC元数据

MARC (Machine-Readable Cataloging, 计算机可读目录)

是一种为描述、储存、交换、处理及检索信息资源而设计的标准, 采用编目条例作为数据建模的规则, 以机读目录格式作为编码方案, 以ISO2709作为交换格式。

MARC元数据

MARC优点

- ◆采用独立的元数据标识系统
- ◆采用纯文本文件
- ◆多年来得到持续的发展与完善
- ◆在世界各图书馆得到普遍应用，积累了大量的元数据

MARC元数据

发展

- ◆1969年，美国国会图书馆发布MARC磁带
- ◆1973年，MARC成为国际标准（ISO2709）
- ◆1977年，UNIMARC（通用MARC格式）出版
- ◆1996年，CNMARC作为行标(WH/T0503-96)开始实施
- ◆1999年，MARC21出版

MARC元数据

◆在IFLA的UNIMARC(1977)、美国的USMARC(1982)、英国的UKMARC(1969)三大格式基础上，各国均纷纷开发出本国的MARC格式

USMARC:

加拿大、西班牙、韩国、拉丁美洲、印尼、匈牙利、挪威.....

UKMARC:

澳大利亚、法国、芬兰、泰国、意大利、丹麦、新加坡、印度、瑞典.....

UNIMARC:

捷克、克罗地亚、日本、南非、中国、俄罗斯.....

MARC元数据

◆各种类型MARC的融合

1990年代之后，越来越多的国家开始放弃原有书目格式，转而使用USMARC以及后来的MARC21，如：澳大利亚(1991)、泰国(1992)、新加坡(1996)、南非(1997)、加拿大(1999)、瑞典(2000)、英国(2002)、印度(2003)、德国(2004)、奥地利(2004)、克罗地亚(2006)和芬兰(2008)

从1994年到2002年，经过多年的努力，英、美、加三国完成了从其国内格式向MARC21统一的进程，2002年，英国签署协议，完全使用MARC21格式。

MARC元数据

◆与其他数据格式的融合

2002年，LC推出MARCXML（适用于MARC21）；

丹麦国家图书馆开发MarcXchange（适用于各种MARC），2008年，成为国际标准ISO25557。

MARC元数据

术语和定义

- 1. 字段：**由字段标识符标识的被定义的特定字符串，可包含一个或多个子字段。
- 2. 子字段：**字段内所定义的数据单位。
- 3. 变长字段：**长度不定的字段，它包含一个或多个数据元素或子字段。
- 4. 定长子字段：**长度固定的子字段。可出现在定长字段，如100字段的@a子字段，也可出现在变长字段，如200字段的@z子字段。

MARC元数据

术语和定义

- 5.内容标识符：包括字段标识符、字段指示符、子字段标识符。
- 6.字段标识符：用于标记字段的一组三位数字符号，也称字段号。
- 7.子字段标识符：由两个字符组成，用以标识变长字段中的不同子字段。第一个字符为ISO2709中规定的专用符号，称为子字段分隔符，第二个字符为字母或数字，称为子字段代码。
- 8.字段指示符：与变长字段联用的字符（数字或字母），为字段内容、记录中该字段与其他字段的相互关系，或某些数据处理时所需操作而提供的附加信息。

MARC概述

指示符：

无指示符；

指示符：空（未定义）

指示符：附注指示符

- 0 不作附注 1 作附注

指示符：题名检索意义指示符

- 0 题名无检索意义 1 题名有检索意义

指示符：题名形式指示符

- 0 与检索点形式不同 1 无检索点形式
- 2 与检索点形式相同

.....

MARC元数据

必备字段：

记录头标 001记录标识号 100通用处理数据
200题名与责任说明 801记录来源

特殊文献类型必备字段：

文字类文献记录：101文献语种

测绘制图资料记录：120字段 123字段 206字段

电子资源记录：230字段 304字段

乐谱等类文献记录：125字段

拓片资料记录：191字段

MARC元数据

CNMARC主要内容

头标区：包含根据GB/T2901的规定所提供的对记录进行处理时所需的通用信息。

0—标识块：包含用以标识记录或标识在编文献的号码。

1—编码信息块：包含编码数据元素。

2—著录信息块：包含除“附注项”和“标准号与获得方式项”以外的其他ISBD规定的著录项目。

3—附注块：包含以自由行文方式对著录项目或检索点做进一步陈述的附注信息，可涉及文献或其内容的物理组成的各个方面。

MARC元数据

4—款目连接块：两种连接技术：嵌入字段技术、标准子字段技术。

5—相关题名块：为编目需要，以特定题名标识那些拥有多个题名的同一作品。

6—主题分析块：包含按照词语或符号的不同体系构成的主题数据。

7—知识责任块：包含需要建立检索点的知识责任。

8—国际使用块：包含国际上一致约定的不适合在0-7功能块处理的字段。

MARC元数据

头标区

定长字段（24个字符）

无字段标识符、字段指示符、子字段

数据元素：

记录长度 0-4

记录状态 5

执行代码 6-9

指示符长度 10

子字段标识符长度 11

数据基地址 12-16

记录附加定义 17-19

地址目次区项目结构 20-23

MARC元数据

头标区

例：018240am2#2200457###450#

（丛书《叶永烈自选三部曲》中的《名人悲欢录》一书的新记录）

专著 低层次记录

记录长度 曾发行较高记录 文字资料印刷品

```
graph TD; H[018240am2#2200457###450#] --> L1[记录长度]; H --> L2[曾发行较高记录]; H --> L3[文字资料印刷品]; H --> L4[专著]; H --> L5[低层次记录];
```

MARC元数据

0—标识块

001 记录标识号

必备，不可重复

无字段指示符、子字段

可自行定义

MARC元数据

0—标识块

010国际标准书号 (ISBN)

子字段表

@a 国际标准书号

@b 限定

@d 获得方式和/或价格

@z 错误的国际标准书号

MARC元数据

0—标识块

010国际标准书号 (ISBN)

例：

010###@a7-118-00249-1@b精装@dCNY55.00

010###@a0-306-35050-5@b全集@dCNY97.29

010###@dCNY3.50

010###@a0-915408-16-2@b模版@d无定价

010###@a0-11-884094-0@z0-11-884094-X

MARC元数据

0—标识块

017 其他标准号

指示符：

指示符1：标准编号类型指示符

7 @2子字段指明来源的标准编号或代码

8 未指明类型的标准编号

指示符2：差异指示符

0 未提供信息

1 无差异

2 有差异

@2 编号或代码的来源。

指明代码类型。不可重复。

MARC元数据

1—编码信息块

100 通用处理数据

必备，不可重复，定长字段

数据元素：

入档时间（必备） 0-7

出版时间类型 8

出版时间1 9-12

出版时间2 13-16

阅读对象代码 17-19

编目语种代码（必备） 22-24

字符集（必备） 26-29

题名文字代码 34-35

MARC元数据

100 通用处理数据

编目语种为汉语 题名文字为广义中文

例：100##@a20000622d1999####em#yochiyo110####ea

一次出全的专著 阅读对象为青年、普通成人 非政府出版物

MARC元数据

1—编码信息块

101 文献语种

只要在编文献有语言文字，则本字段为必备，不可重复。

数据元素：

@a 正文、声道等语种

@b 中间语种（作品非译自原著）

@c 原著语种

@d 提要语种

@e 与正文语种不同的目次页语种

@f 与正文语种不同的题名页语种

@g 与正文、声道的第一语种不同的正题名语种

@h 歌词等的语种

@i 附件语种（非文摘、提要或歌词）

@j 字幕语种（与配音语种不同时）

MARC元数据

1—编码信息块

105-194 各种类型文献的编码数据字段

MARC元数据

2—著录信息块

200 题名与责任说明

必备，不可重复。

数据元素：

@a 正题名

@b 一般资料标识

@c 其他责任者的正题名

@d 并列正题名

@e 其他题名信息

@f 第一责任说明

@g 其他责任说明

@h 分辑（册）、章节号

@i 分辑（册）、章节名

@v 卷标识

@5 使用本字段的机构

@9 正题名汉语拼音

MARC概述

2—著录信息块

200 题名与责任说明

例：

2001# @a 莎士比亚全集 @g sha shi bi ya quan ji @h 第七卷 @i 历史剧 @h 卷一 @b 专著 @f (英)威廉·莎士比亚著 @g 方平主编 @g 屠岸，方平，吴兴华译

2001# @a 罗密欧与朱丽叶 @g luo mi ou yu zhu li ye @b 专著 @d Romeo and Juliet @a 哈姆莱特 @d Hamlet @f (英)莎士比亚著 @g 朱生豪译 @z eng

2001# @a 千家诗 @b 专著 @f 谢枋得选编 @g 郭明志译注 @c 古诗源 @f 沈德潜编 @g 孟庆祥，孟繁红，孟繁翠译注

2001# @a 红楼梦，又名，石头记 @b 专著 @e 程丁插图全本 @f (清)曹雪芹，(清)高鹗著 @g (清)改琦插图 @g 王子风点校

MARC元数据

2—著录信息块

225 丛编项

指示符

指示符1：题名形式指示符

0 与检索点形式不同

1 无检索点形式

2 与检索点形式相同

指示符2：空

MARC元数据

2—著录信息块

225 丛编项

例：

2001#@a巫术@f(法)塞尔韦耶著@g管震湖译

2250#@a《我知道什么？》丛书

461#o@12001#@a我知道什么？

MARC元数据

3—附注块

可涉及文献或其内容的物理组成的各个方面，自由行文。

可遵循ISBD关于附注内容和形式方面的规定，包括标识符的处理。

MARC元数据

3—附注块

327 内容附注

指示符

指示符1：完整程度指示符

- 0 内容附注不完整

- 1 内容附注完整

指示符2：结构指示符

- # 非结构式附注

- 1 结构式附注

MARC元数据

3—附注块

327 内容附注

例：

2001# @a二十四史@h第一卷@i史记 汉书 后汉书 三国志

3271# @a史记/（汉）司马迁撰；梁艺等点校；@a汉书/（东汉）班固撰.....

32711@a序言

32711@a第一章 目标管理之重要性@p1@b1是否从事有效的管理@p1@c(1)管理者的职务是解决问题@p1@c(2)研究问题的性质@p4.....

MARC元数据

4—款目连接块

丛编、补编等：如合订

先前款目：如继承

后继款目：如由……继承

其他版本：如译自、复制自

层级：如总集、分集

其他关系：如被评论作品

MARC元数据

4—款目连接块

两种连接技术：1嵌入字段技术，2标准子字段技术。

嵌入字段技术：

422#1@12001#@a World of knowledge

标准子字段技术：

422#1@t World of knowledge

MARC元数据

4—款目连接块

例：423合订、合刊

嵌入字段技术：

2001#@a文章辨体序说@f（明）吴纳著@c文体明辨序说@f
（明）徐师曾著

423#1@12001#@a文体明辨序说

标准子字段技术：

423#1@t文体明辨序说@a徐师曾

MARC元数据

4—款目连接块

例：461总集

嵌入字段技术：

2001#@a当代经济学译库

2252#@a当代经济学系列丛书

461#1@12001#@a当代经济学系列丛书

标准子字段技术：

461#o@t当代经济学系列丛书

MARC元数据

5—相关题名块

统一题名

不同题名：如并列正题名、封面题名

其他相关题名：如曾用题名（连续出版物）、编目员补充的翻译题名

MARC元数据

6—主题分析块

606论题名称主题

@a款目要素

@j形式复分

@x论题复分

@y地理复分

@z年代复分

@2系统代码

@3规范记录号

MARC元数据

6—主题分析块

6o6论题名称主题

例：

6o6o#@a机械工业@x工业企业@y中国@j手册@2ct

6oo#o@c(宋)@a苏洵@c湄州@f(1009-1066)@x年谱@2ct

6100#@a公共图书馆@a人力资源开发@a培训@a职业生涯管理

MARC元数据

7—知识责任块

701个人名称——等同知识责任

@a款目要素

@b名称的其余部分

@c名称附加

@d罗马数字

@f年代

@g名字首字母的展开形式

@p任职机构/地址

@3规范记录号

@4关系词代码（责任方式）

MARC元数据

7—知识责任块

701个人名称——等同知识责任

例：

701#0@c(日)@a小林规威@4著

2001#@a鲁迅杂文书信@b专著@f周树人著

701#0@a鲁迅@f(1881~1936)@4著

MARC元数据

综合著录与分析著录

无总题名文献：

基本著录：

例1：

2001# @a 罗密欧与朱丽叶 @b 专著 @d Romeo and Juliet @a
哈姆莱特 @d Hamlet @f (英)莎士比亚著 @g 朱生豪译 @z eng

例2：

2001# @a 千家诗 @b 专著 @f 谢枋得选编 @g 郭明志译注 @c
古诗源 @f 沈德潜编 @g 孟庆祥，孟繁红，孟繁翠译注

MARC元数据

分析著录：

基本著录——

2001# @a从地球到月球@b专著@a 烽火岛@f (法)儒勒·凡尔纳著
@g 叁壹编译

分析著录——

2001# @a烽火岛@b专著@f (法)儒勒·凡尔纳著@g叁壹编译

MARC元数据

丛编与多卷书：

综合著录：

以共同题名为正题名，各单独部分的信息著录在内容附注。

分析著录：

编制分析款目，以部分题名著录作为正题名，共同题名著录在丛编项。

MARC元数据

多载体配套文献

有主次之分：

著录主要部分，其余载体作为附件著录于载体形态项，或在附注项说明

无主次之分：

一般文献类型标识：多载体

载体形态项：可著录在一个215字段中，也可重复215字段分别著录不同载体。