数字资源组织的元数据方法

Metadata Approaches for Digital Resources

刘炜
上海图书馆数字图书馆研究所
wliu@libnet.sh.cn

摘要：本文简要分析了元数据方法作为数字资源描述和组织的一般性方法逐渐发展成熟的历史过程，认为元数据方法与传统图书馆的编目方法有许多相通之处，图书馆学的许多领域知识，如目录学、分类与主题方法、编目实践等为元数据方法提供了坚实的基础，结合近年来计算机领域语义万维网的研究开发进展，如知识本体的应用等，有望为数字资源的组织提供一整套方法论体系。文章最后分析了结合知识本体的元数据方法对于数字资源组织的作用和意义。

Abstract: Metadata as a universal description approach for digital resources can be thought of an upgrade methodology of bibliographic cataloguing. The author argue that a lot of expertise by library and information science, such as classification, thesaurus and cataloguing, can be valuable input to metadata approaches for the description and organizing of digital resources. It is becoming a formal methodology of digital cataloguing with the help of Ontologies.

主题词：元数据方法 数字资源 资源组织 资源描述 知识本体 数字图书馆 编目 分类 主题
Keywords: metadata, digital resources, organizing, description, ontology, digital library, cataloguing, classification, thesaurus

元数据是关于数据的数据。这个定义很容易使人们进入到一种无限的递归中去：任何数据都是关于存在的描述，如果把这种存在表达为一种数据，任何数据都是元数据。同时任何元数据都是数据。

进入信息时代，我们人类对于世界的感知越来越多地借助计算机，甚至整个世界表达为数字化信息，再呈现给我们。因特网自不必多说，电影电视是数字的，照片是数字的，广播是数字的，报纸杂志也是数字传送、以点阵印刷出来的，我们对我们目力所及之外的所有感知的事实和观念，几乎都是借助于数字传递过来的，我们与知识有关的一切生产、传递、消费活动，也大多转为数字方式进行，以至于《黑客帝国》（Matrix）之类的电影宣扬了一种虚无的怀疑哲学，人类对世界的感知是真实的存在吗[1]？

好在我们不需要谈论哲学，技术比哲学更实在和更容易些。本文所探讨主题是在上述背景下展开的。一个数字化的世界，充满了各种各样的描述信息，“元数据”无所不在。元数据方法已不再是某些领域特定的专门技术，如图书馆对图书的描述，档案馆对档案的描述，而更具有一般性，成为知识和信息组织的一般方法。

什么是“元数据方法”

正如人们普遍承认的，MARC 是一种元数据，然而“元数据”(metadata)这一名词的产生至少不比 MARC 早[2]。Metadata 来自于数据库领域，一个一定规模的数据仓库必须有一个

[1] 电影 Matrix 中被称为“锡安”(Zion)的人类世界，居然是由母体(Matrix)设计出的！

[2] MARC 创立于 70 年代末，这时代表计算机技术一大进展的关系型数据库还在襁褓之中，尚未得到大规模应用，更不用说数据仓库技术了。

数据字典对各种数据的属性名称进行规定，否则数据就失去了一致性的含义，数据字典里所定义的属性名就叫作 Metadata。由此可见，早期“元数据”与“MARC”这两个概念都局限于狭小的应用领域，而且 MARC 由于定义了详细的磁带交换格式而成为一个行业标准得到了坚实的应用，而“元数据”则由于是一个一般意义上的概念却发展成为一种资源描述方法的总称。

从元数据的定义来看，书本式目录、索引卡片甚至财产帐本都是元数据，甚至可以追溯到刘向刘歆古代目录学的研究对象。然而从实际意义上来看把元数据的研究和应用范围局限在“机读”领域更合适些（也即“数字化”信息），正是因为有了计算机处理信息的强大能力，信息描述及其标准化才成为一个普遍的需求，元数据方法才能够逐渐总结出一套“技能”或者“技巧”应用于实践。从这个观点来看，MARC 以其功能设计的完备性几乎可以称得上元数据方法的鼻祖了。也正因为此，图书馆编目的技能技巧能够为元数据方法提供一个坚实的基础和完美的起点。

尽管计算机信息处理的话语权并不在图书馆行业，然而 DC 元数据的成功使人们看到了图书馆学长期积累起来的专业知识的宝贵价值。对于计算机科学来说，虽然计算机被冠以貌似具有高等智慧的“电脑”之名，然而几十年来大多数成果都只是处理“符号”和“数据”而已，甚至连“信息”都谈不上[3]，“人工智能”并没有取得预期的进展，近年来人们认识到借助于对互联网上信息及关联关系的“语义化”机制，能够发展起一套机器理解和处理语义的“语义万维网”，于是计算机科学才开始关注“语义”处理，才开始研究“语义”的编码、传递和操作的规范和方法，于是就有了元数据方法的中兴。

综上所述，元数据方法可以定义为数字资源的描述方法。数字资源的描述与其如何应用和应用目的是密切联系的，因此元数据方法与数字资源的应用环境和应用领域直接相关，元数据方法旨在提供数字信息资源组织的方法论支持，也即回答“如何进行数字资源组织”这个问题。

元数据方法的内容

数字资源组织有三方面的制约因素：数字资源的属性特征，数字资源的应用环境和数字资源的应用目的。元数据方法并不解决数字资源组织的所有问题，而只解决数字资源的描述问题，然而数字资源的描述与所有这三个方面都有直接关系。

目前发展较为成熟的是关于数字资源属性特征的描述，例如许多数字图书馆都采用基于 DC、按照“应用纲要”的要求适当扩展的元数据元素作为其资源描述的基础。这其中有一整套“方法”已成为一系列的标准规范或“最佳实践”，例如除了 DCMES[4]已成为 ISO/IEC 和 NISO 等国际国家标准之外，“元数据应用纲要”也是一个欧洲标准。就具体的方法来说，有关参考模型的建立、属性元素的选取、扩展和限定规则、著录规则、编码方案、工作流程（参见图 1）、注册体系等等，都正在形成一定的可资参照的规范（具体的元数据方法所包含的文档可参考表 1）。

[3] 一般认为，“符号”、“数据”、“信息”、“知识”、“智慧”这五个概念具有一种递进关系，后者基于前者而高于前者，符号不具有具体含义，数据不具有相关语境，信息不包含价值判断，知识的预见性有赖于智慧等等。然而他们之间的界限并不是非常明确。

[4] 即 Dublin Core Metadata Element Set，包含 15 个基本元素的都柏林核心元数据元素集合。

图 1：元数据方法一般设计流程图示

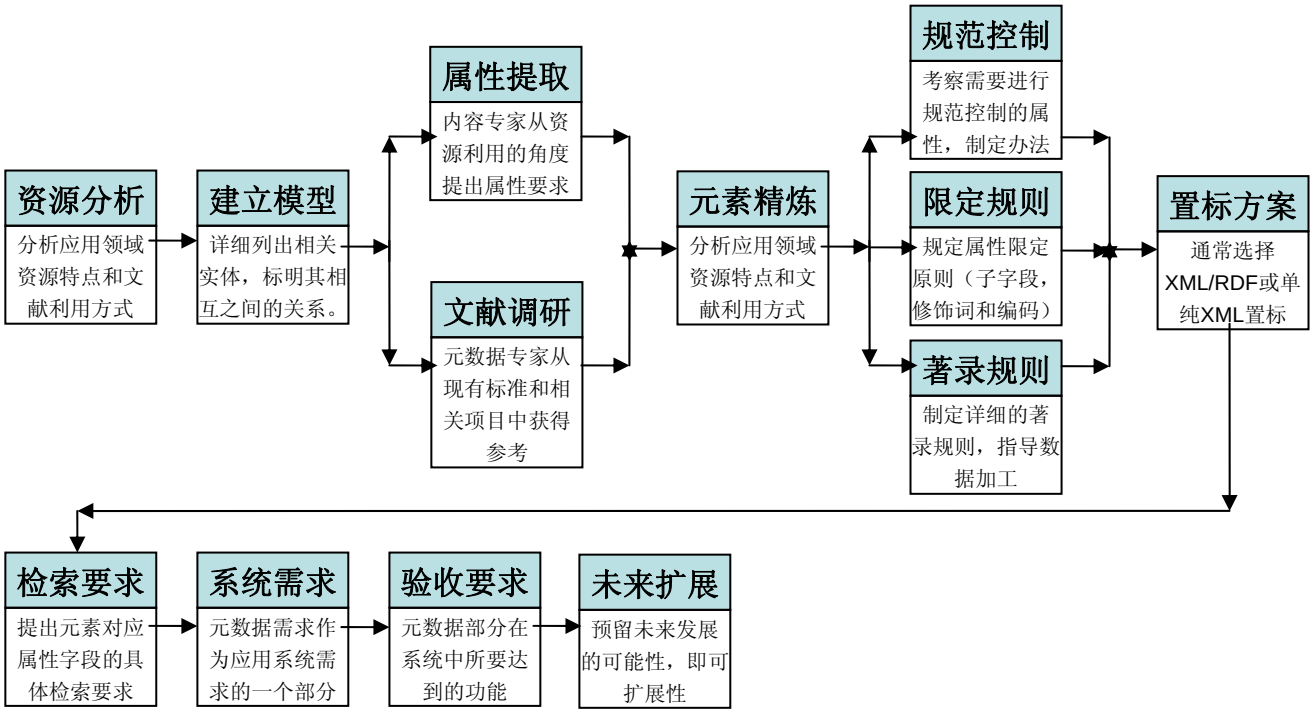


表 1：元数据方法所需一般文档列表

资源分析文档：	定义资源类型，确定资源类型的内涵和外延，确定著录级别和著录单位，属性提取，确定著录对象之间以及属性之间的关系，属性的检索要求，核心属性的支持
元数据元素集及定义：	按照规范格式定义属性元素，定义限定和扩展规则，定义子元素和编码体系，定义元素之间的关系
著录规则：	元数据方案应用于具体资源类型著录时的细节描述
置标方案：	采用 RDF/XML 或其他置标的规则，命名域规定，标签定义，编码规则
规范控制流程：	规范档的建立、维护、更新、应用流程和规则
元数据著录系统需求：	开发具体的著录系统所要求的通用和专用需求规范

然而仅仅是这些还是不够的。作为一种普遍的“元数据方法”还有许多因素要总结和考虑。目前的元数据方法还局限于“数字图书馆”应用领域，以因特网为应用环境，改善数字资源的存取为应用目的。仅就这个领域来说，元数据的应用流程较为注重数字资源内容的属性描述，而对于结构性元数据、管理性元数据、保存性元数据等研究的不多，对于数字资源形式特征、功能特征的提取和描述缺乏方法，对于属性约束和语法信息的描述没有规范，对于语用信息的提取更加缺乏方法。而且不同的数字资源应用环境和应用目的会对描述提出不同的要求，带来了更大的挑战。例如在 Web 服务的技术架构中，电子商务的应用对于元数据方法更是提出了完全不同的要求，传统的基于信息内容的描述方法可能完全没有了用武之地。

数字资源描述是为了实现数字资源的有序化，也即数字资源的组织。元数据方法只是达到有序化的一种方法，计算机科学一直在孜孜不倦地寻求资源组织的内部规律，以求更少地人工干预、更多地自动实现数字资源的有序组织。现在一般认为少量地人工干预能够大大提

高信息组织的效率和人性化“亲和力”，元数据方法也必须考虑最大程度地利用信息技术所提供的可能性，以较少的人工获得最好的效果。

元数据方法与编目

从图书馆学角度来看，元数据方法实际上是对数字资源进行编目的方法。分类法与主题法也是常用的信息资源描述和组织方法，对于数字资源同样有用，但是传统上这两者只是编目的一部分内容，涉及文献的内容属性，而不管知识产权属性、形态特征、管理属性等其他属性，编目则需要根据著录规则著录资源的完整属性（属性集）。对于传统的编目来说，编目规则（AACR2）就是文献著录的完整元数据规范，MARC 就是一种元数据元素集的定义规范，ISO2709 就是一种用于交换的元数据交换。现在 MARC 也发布了自己的基于 XML 交换格式，可以不用 2709 格式，只用 MARC 中的元素集。因为 MARC 本身也是一种元数据标准，因此应用其它元数据方案，例如 DC，方法完全一样：由著录规则、编码方式等等。

在描述对象方面，元数据的描述对象超出了传统的"文献"范畴。根据国标定义，文献是记录有知识的一切载体。按照 DCMI[5]和 W3C[6]的解释，元数据的描述对象是“资源(resource)”，而“资源是具有标识的任何东西”。也就是说，世间万物，只要人能够识别出来的东西，给它一个标识（最常用的标识就是名称），它就成了"资源"，就是元数据可以描述的对象。当然实际上 DCMI 和 W3C 强调他们的"资源"只是在互联网上，由命名域给出标识——URI 的信息资源。

不仅元数据描述的对象远远超出了文献，元数据描述的目的也非常广泛，元数据是语义万维网基石，而语义万维网将是未来一切网上信息活动的一个基本平台，包括电子商务、电子政务、电子教育、电子医疗等各类丰富多彩的应用。当然不管什么目的，元数据本身都只是属性信息，都是为了信息资源的查找 Find、标识 Identify、选择 Select 和获取 Obtain（FRBR 报告中[7]对书目信息功能的定义）的目的，也就是信息的组织。

数字资源组织体系

元数据方法不仅应用于信息系统建立过程中对数字资源的描述和处理方面，而且应用于数字资源组织体系的完整过程。

完整的数字资源组织体系（可以类比于传统的情报检索系统，所不同的地方在于传统的情报检索系统只是索引文摘等二次文献数据库，而这里多为全文数字资源）应该包括以下四个方面：

1. 对资源内容的处理，即数字资源进行结构化描述（元数据著录），按照不同的属性进行有序化组织、索引、链接、建库、存储等；这个过程是资源组织的最主要的过程，传统上这个过程就是信息资源的组织的全部。其中主要的内容组织方法和特征见下表所示。
2. 对用户使用习惯/知识背景(user profile)的处理，例如用户的定制和配置信息、使用偏好、相关反馈的统计信息等）。
3. 对提问的处理，包括语法转换，交互修正，提问分发，规范后控等。
4. 对检索结果的处理，包括剔重以及根据查询结果对于用户的重要性排序等。

包含这四方面的完整数字资源组织体系图示参见图 2。

表 2：数字资源组织基本方法的比较

---

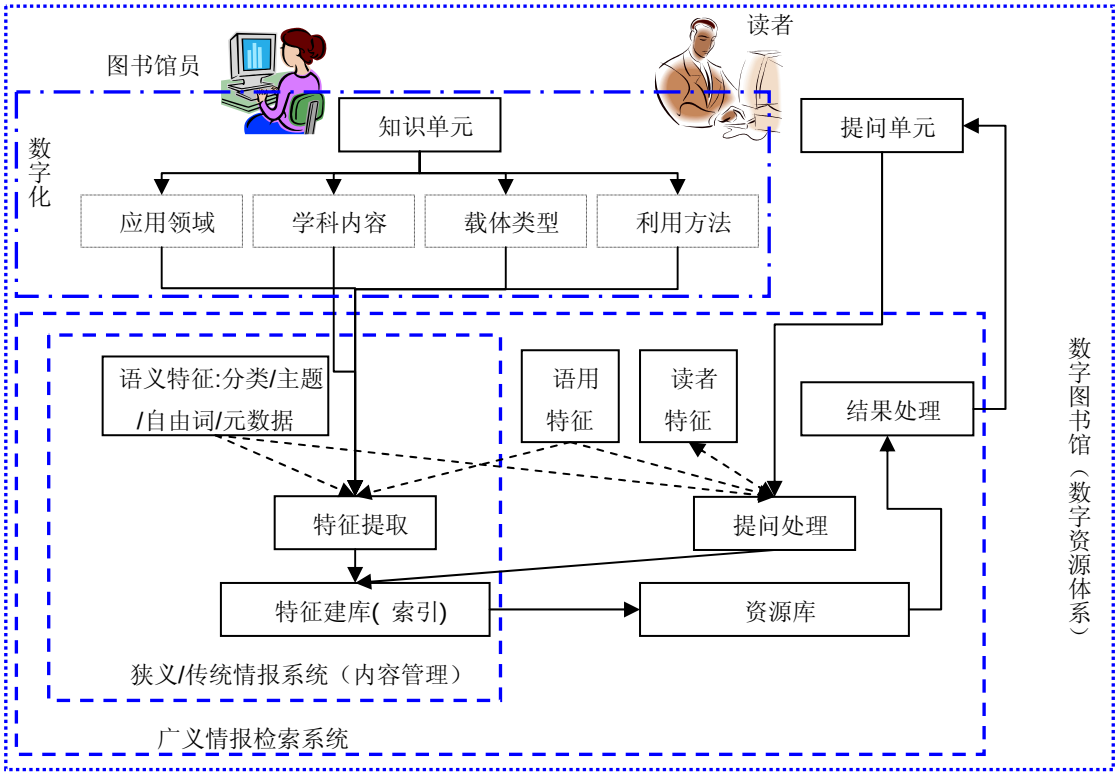
[5] 即 Dublin Core Metadata Initiative 的简称，都柏林核心元数据计划。

[6] 即 World Wide Web Consortium 的简称，万维网协会，负责制定、维护和推广应用与万维网有关的标准规范的组织。

[7] 即 Functional Requirements for Bibliographical Record，书目记录的功能需求，国际图联(IFLA)在一项研究中提出的一种书目信息系统的参考模型。

描述方法	元数据形式	适用资源对象
分类法	一个或几个类号，对于整体内容进行宏观概括	书/论文
主题法	数个主题词或关键词，经过或不经过规范，来自资源内容	论文/篇章/段落/句子
分类主题（知识地图）	代表内容的类目，来自资源内容	书/论文/篇章/段落/句子
全文索引	来自对全部内容（禁用词除外）进行切词、索引、排序	全文（粒度到词汇）
自动分类	矢量/计算整个内容与标准类的距离	整个库/全文
语用信息索引	计算所有词与提问词的距离	整个库/全文
链接法	链接地址	任意部分

图 2：数字资源组织体系图示



可以发现上述完整的数字资源组织体系实际上分为资源组织和检索两个互逆的过程。元数据方法用于资源组织过程自不必多说，同样应用于用户特征信息的描述、提问式切分与规范控制以及检索结果的控制处理方面，甚至可能需要对于整个体系模型的过程描述以便与其它信息系统互操作。

元数据方法与知识本体

知识本体（ontology）本来是哲学中的一个概念，近年来越来越多地应用于数字资源的描述、组织与管理，特别是语义万维网领域。人工智能领域经常引用 Gruber 在 1993 年的定义“概念体系的规范”（specification of conceptualization）[8]，1998 年 Studer 等人在这个定义

[8] 见：http://www-ksl.stanford.edu/kst/what-is-an-ontology.html （2004/4/24）

的基础上对于本体的特点给出了一个较为明确的解释：“知识本体是对概念体系的明确的、形式化、可共享的规范说明”。直观地，我们可以把知识本体看成是“领域知识规范的抽象和描述，表达、共享、重用知识的方法。”

元数据提供了数字资源基本的属性描述集合，使资源有了基本的微观结构，但是不同的数字资源大都采用不同的元数据标准，每个资源对象可以基于不同的目的，从不同角度来描述，因此可以有多套属性元素集合。或许随着标准化的进程，DC 元数据等少数元数据格式将占据主导地位，然而永远不可能统一到仅有少数几种格式。许多专业或专门领域仍然会有大量的元数据方案，这些元数据方案可能局限于一个狭小的领域，都是关于这一领域的一种认识和看法的概念体系（规范词表），都可以看成是关于这件事物的一套领域知识，即本体。只有专业的元数据对于专业的应用才是最合适的，与学科外其他领域的互操作性考虑是次要因素。在网络环境下要联接这些“信息孤岛”，必须有某种程度的互操作解决方案，而且最好是标准的解决方案，单纯的元数据方案不能解决数字资源组织体系的异构问题，包括资源采用不同元数据方案所造成的微观结构的异构问题以及资源对象之间存在的复杂的关联关系，这就需要在元数据之上再建立某些机制，来灵活地实现信息系统之间的互操作。因此元数据方法中还应该包括不同元数据方案进行互操作的方法，这就需要用到知识本体。知识本体的本质就是领域知识的共享和重用，标准化和形式化的领域本体能够为信息系统之间的高层互操作提供很好的工具。知识本体在某种程度上可以看成是关于元数据的元数据，或者说是关于元数据的方法论。

分类法与主题法也都可以看成是知识本体，因为他们都是从学科角度，对描述对象进行归纳或解构。一组资源利用分类法或主题法进行元数据标引之后，在学科空间上可以呈现出一个庞大复杂的"语义地图"，采用不同的分类、主题方法可以呈现不同的语义地图。

元数据方法通过知识本体在不同的元数据标准以及信息资源之间建立起一种普遍联系，并使这种联系"机读化"，大大拓展人类处理知识的能力。体系分类法中的体系，主题词法中的概念关系（主要是用代属分参）都反映了知识单元之间学科属性的普遍联系，都是本体需要实现的重要内容，也是图书馆学长期知识沉淀的成果。当然知识本体中还有更广泛、更复杂的关系，例如信息体的生命周期关系（FRBR 就可以看成这样一种关系）、时空关系（GPS 等地理信息系统、可以用来描述家谱文献），甚至历史上很荒谬的各种认识体系，都可以以本体的形式呈现，并用于组织相应的领域知识。

知识本体弥补了元数据的不足，共同组成完整的元数据方法。本体以规范的方法建立起来，可以支持元数据方案之间的翻译、映射、参照、注册等等功能，进行本体之间的信息交换，使计算机能够无障碍地 "懂得"彼此的语言。

知识本体对于元数据方法的贡献可以总结如下：

- 元数据方案不具有普遍适用性。无法克服特殊性与一般性的矛盾，而形式化的知识本体可以提供一种在元数据方案之间自动映射的机制，进而可以通过语义 Web 服务的体系架构进行实现；
- 元数据应用难以实现元数据方案本身的进化，而知识本体可以提供信息系统的其它视图，只需要通过自动或半自动的手段应用新的元数据方案；
- 元数据方案自身难以对不同知识体系、不同“粒度”的资源进行描述，而知识本体正是起到这个作用，从而实现异构资源和系统之间的语义联系；
- 单纯的元数据方案对于数字资源的整个生命周期的描述非常困难，而采用以诸如 FRBR 模型为基础的知识本体，这个问题便迎刃而解，不同生命周期的知识产权属性也非常易于描述；

除此之外，知识本体同时也在一定程度上解决了元数据方案的灵活性和可扩展性问题，以及在资源集合层面的整合的难题。

元数据方法伴随数字资源描述的需求而产生，吸取了图书馆界在信息资源描述与组织方面的成果与经验，结合知识本体等计算机信息处理技术的最新进展，正在成为数字资源组织的一般方法，同时也将对图书馆学的多门相关学科产生积极的影响。

参考文献

1. 张晓林．元数据研究与应用．北京：北京图书馆出版社，2002－05
2. 陈雪华, 陈昭珍, 陈光华．数位图书馆 XML/Metadata 管理系统．台北：文华图书馆管理资讯公司，民国 90[2001]
3. 肖珑，陈凌等. 中文元数据标准框架及其应用. 大学图书馆学报，2001，19（5），p29-35
4. 吴建中等． DC 元数据. 上海：上海科学技术文献出版社，2000
5. 我国数字图书馆标准规范建设之“基本数字对象描述元数据”子项目组. 基本数字对象描述元数据标准，2004 年 5 月
6. C. Lagoze, “Keeping Dublin Core Simple: Cross Domain Discovery or Resource Description?,” D-Lib Magazine, 7 (1), 2001.
7. [Studer, Benjamins, Fensel 1998] Knowledge Engineering: Principles and Methods. Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Data and Knowledge Engineering, 25(102):161-197, 1998.
8. D. Brickley, J. Hunter, and C. Lagoze, “ABC: A Logical Model for Metadata Interoperability,” Harmony Project, Working Paper 1999. http://www.ilrt.bris.ac.uk/discovery/harmony/docs/abc/abc_draft.html
9. Cromwell-Kessler, W. (1998). Crosswalks, metadata mapping and interoperability: What does it all mean? In Murtha Baca (ed.), Introduction to Metadata: Pathways to Digital Information. Los Angeles: Getty Information Institute. pp.19-22
10. D. Bearman, G. Rust, S. Weibel, E. Miller, and J. Trant, “A Common Model to Support Interoperable Metadata. Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities,” D-Lib Magazine, 5 (January 1999), 1999.
11. J. Hunter, “MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains,” Journal of Digital Information, 1 (8), 2001.