DC元数据

吴建中　主编

上海科学技术文献出版社

DC 元数据

吴建中　主编

上海科学技术文献出版社

责任编辑：胡德仁

DC 元数据

吴建中　主编

上海科学技术文献出版社出版发行
（上海市武康路 2 号　邮政编码 200031）
全国新华书店经销
江苏常熟人民印刷厂印刷

开本 850×1168　1/32　印张 7.5　字数 208 000
2000 年 10 月第 1 版　2000 年 10 月第 1 次印刷
印数：1－2 100
ISBN 7-5439-1580-4/G·412
定价：18.00 元

DC元数据

上海图书馆
上海科技情报研究所《元数据》课题研究组

组　长　吴建中
副组长　陈君辉　刘　炜　孙继林
组　员　赵　亮　纪陆恩　庄蕾波
　　何剑威　高柳宾

序　言

我们正在进入一个网络化时代，据美国微软公司 1999 年预测，10 年以后 50% 的阅读材料将是电子读物。尽管印刷型读物仍呈现有增无减的势头，但电子型读物正以几何级数膨胀。由于互联网的大量普及，网页内容以每 12 个月翻一番的速度向前发展。信息技术的飞跃发展不仅将改变人们的阅读习惯，也将给图书馆带来新的挑战。过去，图书馆员依靠自己有效的信息组织技能，可以将以印刷型资料为主体的馆藏资料整理得井然有序，但是当图书馆信息组织的对象不再局限于馆藏，而延伸到所有可获取信息资源，尤其是浩如烟海网上资源的时候，这一技能是否还那么有效呢？互联网现大约有 4800 万个知识来源，如果要给这些网页编目的话，有人估计需要花去全美国编目人员 24 年的时间，按翻一番的速度，第二年就要花 48 年的时间（见 Martin Dillon, Is MARC Dead? http://www.oclc.org/institute/alamarc1.ppt）。也就是说，按照传统的信息组织方式，图书馆是无法跟上时代发展需要的。在电子型资源正逐渐成为信息资源主流的今天，图书馆应该重新调整自己的发展战略，把相当一部分注意力放在电子资源、多媒体资源和网络资源上。

元数据的研究与开发正成为当今信息网络发展的一个热点。元数据，从定义上来讲，是关于数据的数据，或者说是关于数据的结构化数据。传统的图书馆卡片、图书的版权说明、磁盘的标签等都是元数据，MARC（机读目录格式）和 AACR（英美编目条例）也是元数据的格式，但是由于元数据这一词汇概念起源于计算机科学，又是在人们急需解决网络资源无序化的环境下提出来的，所以当

前元数据的研究重点主要还是网络环境下的数据描述和数据管理问题。

传统书目描述方式的局限性

在相当长的一段时期里，MARC 和 AACR 一直是书目数据描述领域的主流工具。从世界范围来看，绝大部分的书目记录都是依据上述方式编制的，只有2%左右的数据采用了其他著录方式，有人估计即使5年以后，也不过翻一番，达到4%。无论是从数据描述的丰富性，还是从数据检索的查准率来看，MARC 和 AACR 都是名列前茅的，现在还没有哪一种元数据格式可以在这两个方面超过它们。如果说图书馆把信息资源的组织和整理仅仅局限于馆藏资源的话，那么现在 MARC 和 AACR 就足以应付了。但是进入数字时代，图书馆将在超越时空的网络环境下工作，那么，原有的数据描述手段就明显地跟不上形势发展的要求了。

不少人已经意识到 MARC 和 AACR 的局限性，几年前就有人提出废除或修改上述书目数据描述方式，但是一方面没有更好的替代方案，另一方面如果废除或修改的话，将不可避免地出现多种格式并存的无序状态，像日本曾经出现多种机读格式一样，这一弯路带来的不利影响不是一年、两年能够消除的。我国也有不少图书馆走过同样的弯路，造成严重的资源浪费和重复劳动，不利于图情事业的发展。

MARC 和 AACR 的局限性主要表现为以下几个方面：

1. 这种描述手段往往只适用于图书馆；

2. MARC 需要在专门的软件系统中使用，而且不太适应互联网的环境；

3. 修订程序相当复杂，而且也非常缓慢；

4. 适用于完整的、静止的信息内容的处理，不易处理动态的多媒体信息；

5. 编制一条机读记录不仅需要经过严格的专业训练，而且需要花一定的时间。

由此可见，在突飞猛进的网络化时代，传统的数据描述方式已经远远跟不上形势发展的要求。

标记语言环境下的元数据的发展

让我们观察一下书目数据表现形式的演化过程。一开始是书本式目录，100多年前卡片目录作为一项重大发明，在图书馆资源的组织和揭示方面发挥了有效的作用。但是到了计算机时代，卡片目录的优越性逐渐消失，取而代之的是计算机可读目录，这种目录描述能力强，检索效果好。但是无论是印刷型目录还是机读型目录，都有各自的不足之处。印刷型目录是供人阅读的，机读目录是供机器阅读的。反过来，机器无法阅读印刷型目录，人难以阅读机读型目录，而人和机器都可阅读的标记语言，如SGML(标准通用标记语言)就解决了这一问题。尽管MARC和AACR也是元数据格式，但目前人们研究较多的元数据，更多地偏重于电子资源和网络资源的应用。

现在网上的信息检索，主要是用Yahoo、Lycos、AltaVista等搜索引擎，这些搜索引擎的工作方式，是通过自动搜索程序来抓取网页信息，然后以自动拆字(词)做索引的方式建立数据库，造成检索效率低，检索结果数量大，而且有用的信息少，尽管这种方式也是用了标记语言，但这是一种在HTML(超链接标记语言)环境下的只注重页面表示形式而不注重内容的元数据，其主要缺点是描述数据结构性能力差、无法深入到语义内容等。

由此，既能解决数据的结构化问题，同时又能克服数据过于烦琐和复杂的新一轮元数据项目便应运而生，如美国联邦地理数据委员会的地理元数据项目FGDC、适应于档案和原稿的EAD以及广泛使用于图书馆界和情报界的DC等。很多元数据项目都基于

XML(可扩展标记语言)环境,克服了 HTML 的显示能力强而结构性描述差等问题,如 XML 本身没有特定的控制标记,其控制标记可以允许设计者在 DTD(文献格式定义)上表明,给予设计者在控制标记和设定属性方面更大的灵活性和自由度,同时又能够保持元数据创建部门的地方特色和个性特点。

但是,由于这些不同的元数据格式虽有相似性,但彼此之间难以兼容,在 W3C(互联网联盟)的授权下,一些元数据研究部门集思广益,制定出符合多种需要、又有灵活性的 RDF(资源描述框架),来支持互联网上各种元数据格式。RDF 是一个与任何特定语法无关的抽象的资料表达模式,用来反映资源(Resource)、属性(Property)和属性值(Value)。XML 与 RDF 结合起来,使得各种元数据的格式都可以出现或运行在同一个界面上,提高了元数据的规范化和互操作性。

DC 的现状与发展趋势

在众多的元数据项目中,DC 在图书馆界或情报界可以说是应用最广、影响最大的一个国际性项目。DC 即都柏林核心元数据集,于 1995 年 3 月由 OCLC 与 NCSA(国家超级计算机应用中心)联合发起,52 位来自图书馆界和电脑网络界的专家共同研究产生,其目的在于建立一套描述网络电子文献的方法,以实现网上信息的辨识、查询和检索。该项目的中心议题是如何用一个简单的元数据记录来描述种类繁多的电子信息,使非图书馆专业人员也能够了解和使用这种著录格式,达到更有效地描述和检索网上资源。

在过去的 6 年里,OCLC 与各有关机构联合举行了 7 次研讨会,每一次研讨会都要推出一些具体的研究成果,这些研究成果或决定往往都冠以会议所在地的名字,由于第 1 次会议在俄亥俄州哥伦布市的都柏林镇举行,会议推出了"核心元数据集",所以该项目称为:"都柏林核心元数据集"。第 4 次会议在澳大利亚的首都

堪培拉举行，专门探讨修饰词问题，所以被称为"堪培拉修饰词"。第8次会议定于2000年10月在加拿大举行，会议将着重讨论DC大家庭各种元数据格式的互操作性问题。

经过6年多的研究与探讨，DC已被翻译成25种语言，其用户遍及世界各地。最近正计划作为美国的国家标准"ANSI/NISOZ 39.85－200x"，交各有关部门听取意见，最后将以投票表决的方式决定其是否成为国家标准。澳大利亚、丹麦、芬兰等国政府已经将DC纳入国家标准中描述电子信息的一个部分，日本、葡萄牙和英国也紧跟其后。可见，DC的影响正在逐步扩大，有望在不久的将来成为各国都能接受的国际标准。

DC由15个基本元素组成，分成三大部分：内容描述部分有题名、主题、说明、来源、语种、关联和覆盖范围；知识产权部分有创建者、出版者、其他责任者和权限；外形描述部分有日期、类型、形式和标识符。与复杂的MARC格式相比，DC只有15个基本元素，较为简单，而且根据DC的可选择原则，可以简化著录项目，只要确保最低限度的7个元素（题名、出版者、形式、类型、标识符、日期和主题）就可以了。上述15个基本元素又称为"简单DC"。

但有些资料是需要详细著录的，为此又推出了"复杂DC"，即引进修饰词的概念，如语言修饰词（Lang）、体系修饰词（Scheme）和子元素修饰词（Subelement），进一步明确元数据的特性。特别是通过体系修饰词，把MARC/AACR的优点和各种已有的分类法、主题词表等控制语言吸收进来，极大地丰富和增强了DC的描述性和权威性。同时在坚持互操作性的原则下，允许各个DC地方版在15个元素的基础上增加新的元素或新的修饰词。

DC不仅具有可选择性和可修饰性，而且具有可重复性和可扩展性的优点。它规定所有元素都是可重复的，解决了多著者或多版本等重复元素的著录问题，同时它又允许资料以地区性规范出现，并保持元数据的一些特性，以便日后有扩充的余地。

为了推广 DC 元数据项目，OCLC 建立了开放性元数据项目平台，即 CORC(联合在线资源目录)，广泛吸引世界各地的图书馆或个人参与创建元数据。上海图书馆于 2000 年 4 月 11 日加入这一项目，据统计，在发起后不到 18 个月的时间里，CORC 已经发展到 500 家成员，建立了 23 万多条元数据记录，从 2000 年 7 月 1 日起进入商业化运作。同时，OCLC 又利用该研究所，通过讲座和研讨会的方式，培养 DC 编目人员。上海图书馆与该研究所于 2000 年 5 月 23～25 日在上海举行了题为"知识管理与元数据"研讨会，该所所长艾力克(Eric Jul)先生和顾问李华伟博士特地来华作演讲。该研究所近期还计划在中国举办多次类似研讨会，以扩大 DC 在中国各地的影响。

元数据研究与开发的前景

MARC 和 AACR 是否会被 DC 取而代之呢？实际上，这种担心是没有必要的。MARC 和 AACR 也在进一步适应新的发展环境，比如 MARC 为适应网络发展的需要，已经在该格式中增加了 538 字段(系统需求和存取注释)、516 字段(计算机文件类型或数据注释)、256 字段(计算机文件特征)以及 856 字段(电子地址和存取)。同时，为了促进 MARC 在网络环境中得到进一步的应用，美国国会图书馆正在研究制定 MARC 的 DTD(文献类型定义)，使得基于国际标准 ISO2709 格式的数据能自动转换到基于 ISO8879 的 SGML 格式上，适用于各类网络软件和浏览器。

与此同时，DC 也在研究如何将 MARC、AACR 以及广泛应用于图书馆和情报所的各类主题词表吸收进来，从 DC 第 4 次研讨会到第 7 次研讨会，都在研究修饰词的问题，并于 2000 年 7 月 21 日正式发布了 DC 修饰词的建议(Recommendation of the Dublin Core Qualifiers)，使得 DC 更为充实，更切合图书情报界的需要，这也是 DC 能够在这么短的时间内得到国际图书馆界和情报界广泛认可

的一个重要原因。这两者一旦结合起来，就会形成一种更加高效、精确的学术性浏览器。

在我国，一些研究机构已经开发出元数据的应用系统，如由中国21世纪议程管理中心、国家科委、国家计委、国家经贸委和中国科学院等共同开发的"中国可持续发展信息共享示范系统"，已经推出地理、海洋、植物、自然灾害等元数据的数据库，并且向社会开放，而在图书馆界，元数据还停留在研究开发阶段。北京大学图书馆、清华大学图书馆、国家图书馆、上海图书馆和广东中山图书馆等正在积极组织力量开发元数据项目，有的已经完成了格式的制定，如清华大学的建筑元数据项目和北京大学的拓片元数据项目已经进入实验阶段，取得了可喜的成果。我认为：各参与单位应携手合作，加快中文元数据的研究与开发步伐。同时我建议：既然DC已经比较成熟，而且在国际上得到广泛的认可和应用，我们可以参照该格式，或者在DC的基础上形成一个适应中文环境的通用元数据格式。

上海图书馆于1998年起开始从事元数据的课题研究，一些研究成果已发表在上海图书馆网页上。2000年6月，上海图书馆为了加大元数据研究与开发的力度，成立了元数据研究课题组，本书就是课题组成立以后的第一项研究成果。

本书是在课题组成员的共同努力下完成的。第一章由刘炜、高柳宾撰写，第二章第一节由纪陆恩、庄蕾波撰写，第二节由高柳宾、刘炜撰写；第三章由赵亮撰写；第四章～第七章由庄蕾波、纪陆恩、高柳宾撰写。本书有三个附录，附录一《CORC系统DC著录实用指南》在经过OCLC研究所同意后，由吴建中、刘炜、庄蕾波、纪陆恩、高柳宾等，根据查尔德雷斯(E.R.Childress)与科尔比(S.Colby)编写的CORC Practice Input Guide for Dublin Core Resource Description一书共同编译而成；附录二《DC元素与修饰词详表注解》，是纪陆恩、庄蕾波根据几个月来在OCLC的CORC系统在线编目

的基础上共同完成；附录三《参考文献》由孙继林、陈君辉、高柳宾汇编。最后，吴建中、刘炜、陈君辉和孙继林对全书进行统稿和审校。由于有关DC研究的中文资料相当缺乏，课题组成员在这段时间里夜以继日，刻苦钻研，积极从网上搜寻最新研究资料，同时结合自己在线编目的实践，克服了时间紧、资料少等诸多困难，终于完成了本书的编撰工作。此外，《元数据》课题组的研究工作从一开始就得到了上海图书馆王鹤鸣书记、马远良馆长以及王世伟、缪其浩、李道林、周德明等领导和专家的关心指导和经费资助，上海图书馆系统网络中心、采编中心、图书馆学情报学研究所以及上海科学技术文献出版社对本书的出版给予了热情支持和帮助，值此付梓之际，特向以上各位表示最诚挚的谢意。同时，本书是在收集大量文献的基础上经过分析研究后完成的，其中参考了许多专家、学者的观点和资料，在此一并表示衷心的感谢。

由于时间仓促，能力有限，书稿中难免有疏漏和欠妥之处，诚望各位专家、学者不吝指正。

吴 建 中

2000年9月5日于上海

目　录

第一章　DC 元数据发展简史

元数据中的一个标准集——都柏林核心元素集(Dublin Core Element Set),简称为都柏林核心,即 DC。由于它具有简练、易于理解、可扩展、能与其他元数据形式进行桥接等特性,能较好地解决网络资源的发现、控制和管理问题,使之成为了一个较好的网络资源描述元数据集,并正在逐步发展成为世界公认的标准。DC 元数据工作组召开的每次会议都有不同的研究重点,并由浅入深、由泛到专地对 DC 理论和应用问题进行商讨和辩论。在讨论的基础上对 DC 进行了一定的补充和修订,使 DC 在结构和功能上逐渐地完善起来并直接促成 DC 新的发展。可以说,DC 的每一次会议都是 DC 发展史上的里程碑。本章通过对 DC7 次会议及其发展历史的介绍,使读者对 DC 这一网络资源描述方面有重要意义的元数据的来龙去脉有一个比较完整的了解。

第一次会议(DC－1)

1995 年 3 月 1～3 日,第一届元数据研讨会在美国俄亥俄州的都柏林镇(Dublin)召开。会议由 OCLC(联机图书馆中心)和 NCSA(美国超级计算应用中心)主持。与会者有来自图书馆界、档案界、人文学界和地理学界的专家,以及 Z39.50 和通用标记语言标准(SGML)方面的专家。大会旨在确定是否只需要一个简单的元数据元素集就能对网上的各种主题资源进行描述,为进一步发展描述电子资源的元数据元素的定义打下基础。

会议目标主要是为了定义一个能被全球所理解和接受的最低

限度的元数据元素集，它能允许作者和信息提供者描述自己的工作，并能为揭示资源提供互操作性。但是核心元素并不能满足特殊用户团体需要的对象描述。

这届研讨会最主要的成果是设定了一个包含 13 个元素的元素集，后来定名为都柏林核心元素集，简称 DC。DC 是在网络环境中，描述文件类对象所需要的最小元数据元素集。而它的结构句法问题则作为一个执行细节没有进行详细说明。

DC－1 所定义的 13 个元素如下（这 13 个元素在以后的 DC 发展中从名称到内容都有了很大的变化）：Subject（主题）、Title（题名）、Author（作者）、Publisher（出版者）、OtherAgent（相关责任者）、Date（日期）、ObjectType（对象类型）、Form（格式）、Identifier（标识符）、Relation（关联）、Source（来源）、Language（语种）、Coverage（覆盖范围）。

英国的 UKOLN（英国图书馆情报网络办公室）的 DESIRE（欧洲研究与教育信息服务系统）项目，专门对现有的多种元数据类型进行了分析和比较，并把它们分为了三个级别：简单格式、结构化格式和复杂格式。

第一级简单格式，包括的是相对来说未经结构化的元数据，特别是指从资源中自动抽取并索引的。这些数据一般是由搜索引擎产生的。如果用户用它们来查询一个已知条目，它们还比较有用。但用户必须对查出的大量资源进行筛选，并且还可能会错过一些潜在相关资源，因为它们没有使用适当的术语进行索引。

第二级结构化格式，允许使用者不必对资源进行检索或联系，就能对资源的潜在用途或重要性进行判断。这些数据已被结构化并支持字段查询。更重要的是这些简单的数据记录能让非专业用户自己来创造，而不需要什么特定学科的知识。描述一般用手工进行，或者用自动抽取的描述来帮助手工编制，DC 就是其中之一。

第三级复杂格式，该格式具有严格的语义规则和完整的信息

描述手段，它有严格的格式规定和详尽的字段，能够精确、完整地描述信息资源。由于结构的复杂性，该格式元数据系统主要是面向专业人员，只有训练有素的专业人员才能有效准确地利用该格式来描述信息。

在网络资源与日俱增的时代，由作者或站点制作的元数据在很多方面将会变得越来越重要，且在各个级别间有一种跨越的趋势。它所提供的记录是为了调和级别一和级别三这两种极端，而提供的一种简单结构的记录。DC 并不是要替代其他的资源描述类型，而是对它们进行补充。DC 能通过扩展或通过对更复杂的记录的链接来增强其功能，并被对应到其他更复杂的记录中去。

美国国会图书馆的 Rebecca Guenther 指出，DC 具有四个优点：①DC 将鼓励作者和出版者以自动资源发现工具，能收集的形式来提供元数据。②它将鼓励包含有元数据元素模块的网络出版工具的创造，从而进一步简化元数据记录的创建工作。③如果有可能的话，DC 生成的记录能作为更详细的编目记录的基础。④如果 DC 成为标准，那么元数据记录就能被各用户团体所了解。

会议还指出了元数据发展的原则：内在性（intrinsicality）、可扩展性（extensibility）、句法独立性（syntax independence）、可选择性（optionality）、可重复性（repeatability）及可修改性（modifiability）。并确定了将来的发展方向：扩展对象类型，扩大功能范围，建立标准的扩展机制及继续优化已有成果等。

总之，第一届会议主要围绕一个简单的资源描述记录的产生展开了讨论，即广为人知的都柏林核心元素集 DC，并最终达成了共识。它可作为一个统一各种网络资源描述模型的基础。

第二次会议（DC－2）

1996 年 4 月 1～3 日由 UKOLN 和 OCLC 在英国的 Warwick（沃

维克)召开了第二届元数据研讨会。它旨在扩大第一届 OCLC/NCSA 元数据研讨会的影响。出席会议的人员有计算机专家、文本标识人员、图书馆专家、美国数字图书馆倡议项目专家、英国 JISC 电子图书馆项目专家,以及欧洲和澳大利亚图书馆的代表,另外还有 MARC 标准制定团体及一些公司的代表。会议的目的是要明确应用元数据来描述网络资源还存在哪些障碍。与会者认为应该在以下诸方面取得进展:即定义应用语法、开发用户指南,明确扩展机制,定义一个可以兼容不同元数据的框架。

这次会议的主要成果是:提出了"Warwick 框架"的元数据结构的概念。这个框架和 Meta Content[MCF]框架的结构和概念,成为资源描述框架 RDF(Resource Description Framework)的一个基础。在 Warwick 框架中,提出一个元数据的容器结构(Container),它可以包含 DC 以及其他一些不同类型的元数据,而 DC 的 13 个元素仍然没有改变。

Warwick 框架

Warwick 是一种用于不同元数据包的集成和互换结构,它为集成和评价元数据集提供了更大的可能性,从而能够实现现有的和将来的数据描述模型。Warwick 框架具有两个方面的重要性。首先,它提供了一个定义和使用各类元数据的结构框架。其次,把 Warwick 框架作为一个环境,它能允许有特定目的的元数据集开发者对自己的工作进行限制和集中,使其他对元数据感兴趣的团体能独立地在满足自己特定需要上取得进展。Warwick 框架是较早提出解决 DC 与其他元数据互操作性问题的概念方案。

这一结构具有下述特征:模块化,可以包含不同类型的元数据对象;扩展性,可以纳入新的元数据类型;分布式,可以掺引外部的元数据对象;嵌入型,可以将元数据对象像看作具有与之联系的元数据结构的信息内容。数据包主要有三类:原始型、间接型、容