《中文拓片影像数据库》的建设

冀亚平

（国家图书馆善本特藏部金石组）

《中文拓片数据库》自2000年6月在北京召开的“中文文献资源共建共享合作会议”第一次会议提出以来，有关各方面在各自的领域内做了不少工作。作为主持这个项目的中国国家图书馆，馆内投入了专门的人力和一定的财力来完善各项相关规则并进行了该数据库的建设工作。我们的工作主要是从以下几个方面进行的。

一、相关规则的确立

为配合《中文拓片数据库》的建设，2000年北京会议特别是2001年4月台湾会议以后，我们即开始对原有的《金石拓片编目规则》进行修订并更名为《中文拓片编目规则》，使其更具合理性、科学性和可操作性。我们还在实验的基础上制定了《中文拓片机读目录格式使用手册》。2000年11月13日，在中国国家图书馆召开了北京地区图书馆界有关专家的座谈会，对规则和手册征求意见，我们根据专家的意见，进行了多次修改后于2001年12月出版发行。还制定了《中文拓片影像加工规则》《数字资源整合规则》《数字资源存储规则》等。

编目是数据库的内容核心，石刻拓片的著录内容历史上一向繁简不一，但是，有几项是必须著录的，如题名、责任者、年代、地点等。在此基础上，我们根据读者的需要，增加了诸如录文、相关文献等项目，共计30余项，供必要时选择使用。

中文拓片机读目录格式的试验是在1999年初开始的。当时，按照馆里的要求，我们采用国内通用的CNMARC格式建立了简体字版的《北京地区石刻拓片书目数据库》6300余条作为第一个实验数据库，2000年又建立了《墓志拓片书目数据库》3400余条。2000年《中文拓片数据库》项目提出后，大家提出要使用一种国际上通行而简便的、适合互联网操作的中文拓片机读目录格式，多数人的意见趋向于Dublin Core(DC)元数据格式。我们仔细研究了中文拓片的特性并对MARC和DC两种元数据格式进行了比较，最终确立了先用MARC格式进行数据加工，然后将其中核心数据转换成DC形式与影像挂接后进行网上发布的原则来构建《中文拓片数据库》。其原因有二：

（一）MARC元数据是专业图书情报机构书目数据的基础，它适用于拓片数据库的建设

上海图书馆馆长吴建中先生对传统书目描述方式曾有过阐述：“在相当长的一段时间里，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。…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。不少人已经认识到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，像日本曾经出现过多种机读格式一样，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”（《DC 元数据》）。从吴建中先生的这段论述可总结出以下三层意思：一是现今世界图书馆的馆藏资料编目多数是 MARC 和 AACR 格式。二是传统的 MARC 和 AACR 格式对现今图书馆的馆藏资料编目仍然适用，但不易处理动态的多媒体信息。三是盲目地废除或修改 MARC 和 AACR 格式可能会带来多格式并存，造成重复劳动和资源浪费。

从中文拓片资料的性质看：首先，拓片资源属于静态馆藏资源中的特殊一类，与动态的网络资源、多媒体信息不同，它不是瞬息万变、无穷增长的信息资源。其次，中文拓片从其制作过程来说，它的撰文、刻石与椎拓过程与中文图书的撰文、制版印刷制作过程是一致的。目前，中文图书在世界范围内广泛采用 MARC 等传统的格式来制作机读数据，那么，中文拓片也可以借鉴其中的成功经验，合理地使用 MARC 这种传统的格式来建立中文拓片数据库。这样做的一个好处是可以将有限量的拓片数据溶入到现已存在的大量的传统的图书数据库体系中共同运行，对于拓片资料的利用和进一步的开发大有好处。从人力资源角度来看，现在中文拓片绝大多数都保存在世界各地的专门的图书情报部门，这些部门的工作人员有着丰富的图书馆专业著录经验和制作 MARC 等传统书目数据的经验。拓片数据库采用 MARC 格式制作，有着比较强的技术保证。制作一条 MARC 格式的拓片数据虽然要比使用 DC 等简便格式多耗费些资料和人力，但从更大范围处着眼，拓片数据库能与有大量的中文图书数据库共同运行所节省的资源和人力相比，还是值得的。因此，我们认为：就有限的中文拓片资源来说，使用相对比较成熟的 MARC 格式来做机读数据，应该是一个比较适当的选择。

（二）网上检索和编目要求 MARC 向 DC 转换

如前所述，MARC 等传统格式不太适应互联网的环境，这也是大家感到不太方便而不愿使用 MARC 的一个重要原因。为了适应互联网环境，我们提出将 MARC 格式的拓片数据核心部分转化成 DC 形式的表单数据与影像挂接后上网提供数据库检索的方案。并于 2001 年香港环太平洋国际论谈及历史文化地图会议时将研制的 1000 余条数据上网公布，效果还是比较好的。对于不习惯使用 MARC 格式进行数据加工的编目单位，可以使用这种简便的 DC 形式的表单来参加互联网下的联机编目。

二、　　中国国家图书馆《中文拓片影像数据库》的建立

中国国家图书馆藏有各类石刻拓片(不含丛帖子目)近 3 万种 13 万余件，占全国现存石刻拓片总量的 60%，其中已编制目录的有 2 万种，已制作出书目数据的有 9700 余种。为了早日在互联网上展示拓片的原貌，并为《中文拓片数据库》做出示范，2000 年下半年我们即开始《中文拓片影像数据库》的建立。该数据库的建立分为二个步骤：

（一）数据制作：

1.按照《中文拓片编目规则》对拓片进行整理编目；

2.按照《中文拓片机读目录格式使用手册》的要求使用 MARC 格式制作拓片的机读数据；

3.将 MARC 格式的数据转换成 DC 形式的机读数据。

（二）影像制作：

1.对加工的拓片进行数字化处理，以数字方式存储，展现拓片原貌，既利于保存，又便于应用。

2.数字化扫描参数。为了保证数字化后的拓片的不同应用，我们按四种技术指标存储数据，即存档数据、高清晰度数据、中分辨率和低分辨率数据。互联网发布图像数据按 150dpi、72dpi 和小图标方式提供使用。

（三）由 MARC 数据向 DC 数据转换，并实现与影像数据挂接

为了实现网上检索查询，在拓片 MARC 数据的基础上进行了拓片元数据的转换，并实现了与拓片影像对象数据的挂接

（四）在互联网上为读者提供检索查询和浏览服务

利用加工软件生成的元数据，我们提供了基于 WEB 的检索、查询与浏览，并实现与拓片的图像数据的挂接。

检索途径：可以从题名，责任者：撰(绘)者、书者、镌者，年代(朝代、年号、年、月、日)，地区(省、县、具体处所)，关键词：包括石刻的形态(摩崖、幢、碣、阙等)、载体类别(甲骨、金属器、石、竹、木、玉、陶、砖、瓦、泥等)、内容(佛经、儒经、道经等)、用途(法帖等)、文种(满文、藏文、拉丁文等)、体裁(诗词、楹联、格言等)、书者(颜真卿行书)朝代、年号等，馆藏索书号等不同的途径查找所需的拓片资料。

排序按原刻立石年月时间顺序显示目录(在数据加工时我们将农历转换为公历，如清乾隆 20 年 2 月 1 日为 17550313)。

该软件提供 MARC 格式的数据输出，并提供拓片元数据的输出接口，各馆可根据本馆的元数据格式进行元数据的输出。

2001 年 2 月，1000 余条中文拓片影像数据首次上网成功。在上网实验取得成功以後，我们又对上网数据的查询功能中存在的问题进行进一步的改进。2001 年我们又完成中文拓片影像数据 4000 余条。现在，大家能在因特网上查询、浏览 5000 余种 7000 余幅中文拓片的影像数据(网址：www.nlc.gov.cn/RubbingImg)。

三、今后工作计划

中国国家图书馆藏有各类拓片数万种 21 万余件，这些拓片内容十分丰富，未来几年，我们将继续以馆藏为基础建立影像数据库，为《中文拓片数据库》的建立打下基础：

1.2002 年度，完成制作 6000 种影像数据的任务。2003 年度计划完成 3500-5000 种。此后，根据经费支持情况，再经过几年的努力，完成剩余的 1 万余种中文石刻拓片数据的制作。

2.准备馆藏法帖、画像、甲骨、青铜器、砖瓦等拓片及印谱影像数据库的建立。

3.在完成建立馆藏拓片影像数据库的基础上，进而与其他国家和地区的图书馆等单位合作，最终实现《中文拓片数据库》的共建共享。

附：

1. 所负责项目简介：《中文拓片数据库》是中文文献资源共建共享之一。以中国国家图书馆所藏拓本为依据建立而成的《中文拓片影像数据库》包含图像和基本数据(繁体字)两部分，首先制作的是碑刻部分。目前，能在因特网上查询、浏览的中文拓片的影像数据有 5000 余种 7000 余幅。
2. 会上发表文章：《<中文拓片数据库>的建设》；报告人：冀亚平
3. 主要联络人：冀亚平
4. 已开发成果的网址：www.nlc.gov.cn/RubbingImg