中国地方志（旧方志）资源库的设计与展望

王荟

（国家图书馆分馆技术服务组）

一、中国地方志（旧方志）资源库建库缘起

地方志，也称为“方志”，“志”就是“记”，是记录、记载、记述的意思。顾名思义，“地方志”就是一个地方从古到今，从自然到社会方方面面情况的总揽。它由地方政府组织专门人员，按照统一体例编写而成，是中国特有的一种文献形式，也是中国国家图书馆具有特色的馆藏。特别是我馆所藏 1949 年以前编纂的地方志（简称旧方志），不论质量还是数量均居海内外各藏书单位之首。对于先人留给我们的这一宝贵文化遗产，我们有责任保护、推介、利用好。采用数字化的形式，整理、加工旧方志资源，是实现这一任务的有效途径。由于旧方志是我馆特色馆藏且不存在版权问题，所以我们准备先制作中国地方志（旧方志）资源库（本文以下简称方志资源库）。

方志资源库采用什么样的结构模式，是建库之初首先遇到的问题。国家图书馆资源库的发展历程伴随着图书馆发展的脚步，也经历了从传统馆藏资源——馆藏资源数字化——数字图书馆资源库发展三部曲。我馆现在已有的数据库，大多是将传统馆藏经过数字化处理生成的。方志资源库如采用原有的思路，即是把志书平行地转换成数字化内容，那我们最终实现的仅是把纸质志书转换成电子志书。就载体形态来说，志书数字化使其发生了质的飞跃，但就内容而言，它仅是形式的变化。而当今信息网络技术的发展，信息网络资源的迅速扩大，越来越影响和改变着人们选择信息、使用信息的行为和方式。人们对知识的需求已不满足于以文献为单位，而是希望通过简单、快捷的方式检索到所需文献中的具体事件、数据、结论等知识单元，同时将所需知识单元和与其相关的信息进行整合，在最短的时间内获取最大的信息量。为了最大限度地开发方志资源，多角度、多侧面地深入揭示方志内涵，依据信息时代人们的信息需求特点和信息网络技术发展所提供的技术保证，达到知识创新的目标，方志资源库的建库模式应定位在数字图书馆的平台上。在此，先论述图书馆数字化和数字图书馆之间的联系与区别以及其它几个基本概念。

二、方志资源库的几个基本概念

1．图书馆数字化：传统图书馆馆藏经过数字化技术处理和加工，为用户提供服务，这个过程就叫图书馆数字化。这方面的工作我们在十几年前就开始了。例如：八十年代后期我们就在 M150 机上编制“国家书目”，九十年代前期我们开始编制的地方志书目数据库及以后编辑的地方志人物传记索引数据库，直到目前我们还在做的“古籍书目”数据库以及我们将要建的地方志资源库中的全文影像库等都属于图书馆资源数字化的范畴。其工作对象和结果仅限于传统资源本身，或者说它只是一种载体形式的转换（由纸制品转化成电子出版物）。

2．数字图书馆：数字图书馆是把传统图书馆的功能由信息的查询和图书资料的借阅扩展到知识服务的新阶段。其中两个关键的技术是信息资源的整合和知识的创新。前者就是要依据统一标准，将相互关联的信息资源重新组合并进行科学的分类和标引，强调重组后的信

息资源的良序化和关联性，而后者突出的是知识的增值与创新。如果把前一过程比做物理学中物质所发生的“物理变化”的话，则知识创新所引发的就是物质的“化学变化”。即通过分解、重组，形成了新的信息知识网络，较原来的传统信息资源在功能、用途等方面都发生了质的变化。对此许多专家、学者都有严谨的描述和解释，在这里就不一一赘述了。通俗一点儿说，它有以下几个特征。

（1）信息资源数字化：数字图书馆内的所有信息资源都经过数字化处理。

（2）服务手段网络化：它借助网络技术、计算机技术和现代通信技术传播知识（例如互联网、卫星传递等），突破了馆舍的时空局限，用户可以随时随地得到所需信息，是不局限于图书馆场馆的。

（3）资源实体虚拟化：它是基于互联网的多维知识网络，突破了传统载体的限制，延伸、拓展了传统图书馆馆藏外延。它不仅提供传统的基于印刷介质的服务，还可通过跨库检索，对数字信息进行重新组合，提供重组后的信息服务。因而，它是对馆藏资源的再开发。例如：“昭陵”和“玄武门之变”分别是景观和事件对象资源，它们都与“李世民”相关联，分别收藏在景观库和事件库中，但重组在一个页面里，以视频、音频、图像等多媒体手段展现在用户面前。

（4）检索方式良序化：依据统一规范即统一的元数据标准，对数字信息资源进行科学的分类和标引，达到对数字信息处理的良序化（相当于书刊采访到馆后先编目），保证了分散的数字资源经重组后提供给用户精确的检索，检索效率很高。这一点有别于一般的网络搜索引擎。网络搜索引擎是通过网络机器人自动搜索并生成相关的著录信息，存入数据库中供检索之用，其检索系统由于采用自动标引，检索后的网上信息还需要人工识别处理，检索效率太低。

（5）信息利用共享化：由于数字化图书馆内的信息资源的加工、发布都依据统一的标准和规范，所以它可以最大限度地实现信息资源的共建共享。

3．元数据：传统图书馆流程中重要的一步工作就是图书、期刊的编目，便于用户查找。数字图书馆中的数字资源同样需要编目。元数据就是为了满足数字资源的编目需要应运而生的，它是数字图书馆编目的新格式，是一种有效的信息资源组织和管理的工具。它具有描述性、结构性、管理性。就描述性而言，我们以前用的卡片目录，现在用的 MARC 格式都属于元数据的范畴。但元数据又比卡片目录、MARC 格式具有更强大的描述能力。而元数据所具有的结构性（揭示资源的内部结构）和管理性（规定运行环境、数字版本、收费情况等）使它能更全面的反映了数字文档的各个方面，为数字资源的保存和利用提供了更有效的工具。

4．资源库：经过专业人员组织、加工、整合而成的符合数字图书馆规范的资源的集合。它具有数字图书馆的基本特征。资源库是完全网络化的，具有强大的检索平台和丰富的检索途径。一般资源库都包括了文字、图片、视频、音频等丰富的多媒体资料，对各种媒体都具有良好的支持。各种专题知识资源库组成了数字图书馆的物质基础。资源库的建设不可能一蹴而就，它要随着时间的推移、知识的不断更新，动态地更新内容。由于资源库的信息资源使用统一的加工、发布标准（例如元数据标准），可吸收各信息资源优势单位参与共建，形

成系列知识库群，使信息资源最大限度地被公众共享。

三、方志资源库的结构和内容

下面介绍“方志资源库”的结构设计。中国数字地方志资源库由一个全文影像库，一个OCR数据库和八个专题子库组成。

全文影像库：就是将国图分馆所藏的1949年前编纂刊行约6000余种地方志书进行全文扫描，即志书的数字化处理，全文影像库内的数字化资源并没有改变原有的信息组织，它只是对纸质志书的载体形式进行了一次平面转移，即将纸质旧志通过数字化处理（扫描），变成可在网上阅读的电子书。在全文影像库用户除了阅读原书，还可以做多幅影像的比较即版本校勘。（最多4幅图像同时显示）。

OCR数据库：OCR是英文Optical Character Recognition的缩写，意为“光学字符识别”，也可简称为文字识别，通俗地说就是计算机认字，是一种文字自动输入方法。它通过扫描仪获取纸张上的文字图像信息，再和计算机配合，经OCR软件将图像数据进行运算分类后，将图像数据转换成计算机内码，并按规定格式存储在文本文件中。它的作用是将全文影像库中的志书影像转换成文本格式再进行切词标引，按设计要求规定标引到志书中的“标目”。这样在OCR数据库中，用户不但可以阅读到旧志的原文，还可以对志书进行全文检索，也可进行精确到“标目”的词组检索。同时在OCR数据库中，用户还可以根据需要进行个性化处理，例如添加标记、注释，选择自己需要的内容进行编辑、复制，对不同版本的影像进行多屏幕比较研究等。

以上两个库内容的外延都没有超出志书提供的内容。而只是将原书载体形式做了转换，以电子图书的形式显现。

八个专题子库：

八个子库的建库原则是以原书内容为基础，按照元数据标准进行规范化处理，多角度、多途径地丰富、补充、扩展原书内容，将原来分散或不完整的方志信息集合起来，形成地方志知识网络。八个专题资源库彼此相连，并且都与全文影像库和OCR库相连接，专题资源库的检索条件可以单独使用，也可以两个检索条件组配，进行复合检索。在专题资源库，用户根据需要，从一个知识点入手检索，就可以方便、快捷地跳转到全文影像库、OCR库或不同的专题资源库，检索到与入口知识点相关的各种信息，从而为用户节省大量的精力和时间，最大限度地为用户提供个性化的服务。

地名资源库

地名资源库的建设分为两步：首先建立志书名称中涉及到的方志地名资源库，其次再逐步扩大地名收录范围，成为更大规模的中国地名资源库。地名库中的规范地名是依据 2002年国家行政区划表及有关规则进行规范处理后的地名。客观地名则取自志书卷端题名。地名异名包括客观地名的又名、别名、俗称、简称等。规范地名与客观地名相互参见，地名异名见客观地名。用户可根据需要选择入口词。该库的基本内容包括：规范地名、客观地名、客观地名的异名、地名隶属关系、地名沿用时间（朝代）、地名简介（沿革、变迁情况）、地名文化（相关人物、事件、景点和插图名称等）、周边地名、所辖地名、影像原文和OCR原文

等。

人物资源库

包括出现在方志人物、选举、职官等篇目中的有传记资料的人物，按照元数据标准进行规范处理，规范人名与又名（别名、笔名、室名等）之间建立相互参照关系，用户可从被传人物的任一名字入口检索所需人物信息。人物资源库基本内容包括：规范人名、又名（别名、笔名、室名、字、号等）、性别、籍贯（出生地）、民族、生卒年、主要活动年代、人物关键词、分类。通过链接相关信息（相关人物、地名、事件、插图、景点、文献等）和影像原文及 OCR 原文，可以多侧面、多角度地描述被传人物。

事件资源库

事件资源库收录了志书大事记篇或杂记中记载的重大事件。内容包括灾祥、战事等。依据元数据标准，对事件资源进行规范著录、标引。事件资源库基本内容包括：事件名称、发生时间、地点、事件简介、分类、关键词、相关信息（相关人物、事件、地名、作品、景点、志书、文献等）、影像原文、OCR 原文和出处等。

作品资源库

该库收录了方志艺文志、经籍志、人物志中记载的诗词、游记、散文、墓铭志等作品，依据元数据标准进行规范著录、标引。它与相关文献资源库的区别在于作品库的内容一定出自志书，所做的补充和扩展都是为了保证志书中作品的完整性而进行的。例如：《武功县志》上有关于骆宾王的记载，而艺文志中他的诗文作品收集的又不全，这时为了保证有关骆宾王诗文作品在作品库中的完整性，可以从其他途径进行补充和完善。但所有补充的内容必须是骆宾王的作品。

作品库内容包括：作品题名、作者、作品出处、出版情况(包括出版者、出版地、出版日期）、现存版本、发表时间或历史时期、关键词、分类、原文（图片资料、视频、音频资料）等、作品出处、提要和制作信息，并增加了相关地名、人物、事件、景点、插图、研究文献、影像原文和 OCR 原文链接，使作品资源库内容更丰富、充实。

插图资源库

插图资源库将志书中的舆地、器物、肖像、景观和营造等类插图，依据元数据标准进行规范著录、标引，建立插图资源库。该库内容包括：插图代码、名称、版框尺寸、出处、关键词、分类、相关链接（包括相关人物、地名、事件、景点、文化民俗、研究文献等）、图像信息（包括图像格式、图像文件大小、尺寸、分辨率和色彩深度等）。用户可通过插图名称、关键词、分类号等途径，检索到志书插图并能自动连接到同一志书的其他插图。还可以通过“原图”、“原文”按钮或输入 URL 网址与全文影像库、OCR 库链接。

景观资源库

该库将方志中记载的名胜景观，依据元数据标准进行规范著录、标引，建立景观名称和它的又名之间的相互参见关系，并在该库中增加了景观图片、视频、音频资料，相关人物、事件、作品、地名和研究文献等信息，使景观内容更丰富、充实。该库内容包括：景点名称、位置、景观介绍（包括文字、图片、音频、视频）、分类、关键词、景观文化（包括相关人

物、事件、作品、地名、研究文献等)、周边景观、下层景观并可以链接到影像原文或 OCR 原文。

目次资源库

该库记载国家图书馆志书收藏信息。依据元数据标准，对志书题名和志书中的卷次篇目进行规范处理。通过该库用户不但可以检索到志书书目，还可以对书中的卷次篇目进行检索。目次资源库的基本内容有：志书名称，目次名，版本项、载体形态、相关信息（地名、人物、事件）等。

相关文献资源库

收录后人对志书、志书版本、志书内容（相关人物、事件、地名等）的评论、研究、考证论文、论著，依元数据标准进行著录、标引。该库基本内容有：文献名称、作者信息（责任描述、工作单位)、原文、分类、关键词、文献发表时间、发表刊报、相关链接（包括相关事件、人物、作品、志书、地名、插图、文献)、影像原文、OCR 原文等。

全文影像库、OCR 数据库和八个专题子库相互关联，互为依托，构成数字方志资源库的主体。

四、方志资源库的建库进程及展望

我们在对馆藏情况（包括方志数量、质量、版本等）进行认真分析、调研后，就方志资源库的结构、规范、收录范围、工作进度、共建共享等问题反复论证、规划，在此基础上，我们推出了方志资源库演示版，并在 2002 年 7 月召开的《北京国际数字化公众信息服务与技术展览会》上进行了展览和演示，得到了与会各级领导和专家及观众的关注和肯定。现在我们已经开始了全文影像库的建设，到 2002 年底，计划完成 330 万页旧志的全文扫描，同时，对 7 月推出的演示系统进行完善，从明年开始着手制作与之配套的 OCR 数据库及地名、人物、事件、作品等八个规范化的专题子库。方志资源库完全建成后，应包括本馆所藏 6000 余种地方志（旧方志）及其相关信息资源。此外，我们还设想待今后条件许可继续扩大收录范围，一方面希望联合全国地方志（旧志）收藏单位，参与我们方志资源库的建设，另一方面将吸收我馆普通古籍中所藏的各种专业志资源，丰富其内容，将其建设成为更广大意义上的中国地方志（旧志）资源库。

通过上面叙述，可以看到作为数字图书馆重要物质基础的方志资源库是一个多维的方志信息资源网络。它的建设是一个非常宏大的工程，也是一件非常艰苦的工作，是需要耗费大量的人力、物力和财力才能做好的，但同时这又是一件造福后人的工程，是一件很有意义的工作，值得我们为它去努力。

国图特色鲜明的馆藏为开展数字方志资源库建设提供了资源保证，相当长时间内我们已经开展起来的二次文献开发和数据库建设为资源库建设积累了宝贵的经验，业已形成的图书馆自动化和服务网络为资源库建设提供了技术和手段保证，更重要的是我们有一支熟悉馆藏、熟悉古籍的专业人员队伍，依靠着这些优势，国家图书馆方志资源库的建设一定会有一个光明的未来。