古籍书目数据库整改工作构想

鲍国强

世纪交替之际，内地的古籍书目数据库从试验阶段进入了推广和实用阶段。在 CNMARC 框架下，国家图书馆于 2001 年 8 月约有 15 万条善本和普通古籍书目数据上网供读者检索，到 2002 年 6 月已达 18 万条，高校图书馆的 CALIS 系统于 2001 年底展开了古籍书目数据的联合编目工作。其他各古籍收藏单位根据自己的馆藏目录电子化计划也做了大量的古籍书目数据制作工作。这些古籍书目数据库或上网运行，或作本地查询，极大地方便了广大读者快速检索相关的馆藏古籍。

但毋庸讳言，由于古籍本身的复杂性、机读格式的不断完善和古籍书目数据制作工作起步早晚不齐等原因，这些大量的古籍书目数据在格式和内容等方面还存在着不少的问题，使它们难以发挥更大的作用。再者，随着互联网文献检索技术的日益提高和普及，古籍元数据的网络应用也提到了议事日程。古籍的机读目录数据与 Dublin Core（都柏林核心集）元数据是可以互相转换的。只有提高古籍机读目录数据的质量，才能保证古籍元数据的质量。

目前，国图等馆为了尽早实现馆藏古籍目录电子化而进行相对简约快速的古籍书目数据制作工作已经进入后期，从机读目录格式和网络检索最新要求出发整改古籍书目数据库，提升古籍书目数据的整体质量，已经成为需要切实研究的课题。此课题的研究对于将要和正在进行古籍书目数据制作工作的单位也不无借鉴意义。本文试结合笔者从事古籍机读格式编制和书目数据制作的工作体会，阐述网上古籍书目数据库的整改工作要点，供同业者参考。

一、修订古籍机读目录格式

从格式方面整改古籍书目数据库的依据是标准的古籍机读目录格式。2001 年 10 月由北京图书馆出版社出版的《汉语文古籍机读目录格式使用手册》是根据 1996 年 2 月中华人民共和国文化部发布的行业标准《WH/T 0503—96 中国机读目录格式》和 2001 年 3 月科学技术文献出版社出版的《中国机读目录格式使用手册（修订版）》制订的。目前，国家标准的《中国机读目录格式》正在制订当中，新的《中国文献编目规则》（含古籍分则）也在修订。它们的正式发布和实施，将进一步推动古籍机读目录格式的修订工作。

修订古籍机读目录格式主要应考虑以下三方面内容：

1．重新确定汉语文古籍的有关编码字段。国家标准《中国机读目录格式》（征求意见稿）

鲍国强，国家图书馆研究馆员。

认为，140 字段（古籍——一般性数据）和 141 字段（古籍——复本特征）适合外国古籍（以圣经及相关文献为主体的古籍文献），不能反映我国以儒家文化为主体的古籍传统特征，故另自定义 193 字段（中国古籍——一般性数据）和 194 字段（中国古籍——藏本特征数据）来记录中国古籍相应的编码数据。这些 193 和 194 字段就是在古籍机读目录格式中应该重新确立的汉语文古籍编码字段。

2. 明确古籍类目在相应字段的表达格式。这里主要有两个问题：一是古籍类目使用 686 还是 696 字段？686 字段应该用于已在国际标准组织注册的其他分类法分类号。四库分类法等我国古籍分类法分类号（类目）应该使用 696 字段。二是四库分类法的次级类目在 696 字段中如何体现？如：集部总集类。“集部”入@a 子字段，没有分歧。“总集类”如何处理？目前见到的主要做法有三：（1）“总集类”前加空格，与“集部”一起入@a 子字段。（2）重复@a 子字段，著录“总集类”。（3）“总集类”入@c 子字段（分类复分）。这是需要明确的问题。

3. 如何在古籍机读目录格式中贯彻“完整本著录原则”？“完整本著录原则”是 UNMARC 为了适应网络联合目录和本地版本品种目录需求提出的，它主要应用于古籍机读目录格式（因为普通图书的残本问题不突出）。所谓“完整本著录原则”，简言之，就是不管手头所编的古籍是全本还是残本，要把它完整本的信息记录在著录正文，流传过程中出现的特征（残本、印章和批校题跋等）及其馆藏信息均入复本字段。这样做可以解决同一版本的一条古籍记录中本馆或其他馆补复本的问题，但这个原则与我国的传统做法区别很大，我们原来是流传过程中出现的特征稍有不同便另做一条记录。另外，原来的古籍机读目录格式已经移植一些反映“完整本著录原则”的做法，如使用 316 等复本字段和其中的$5 子字段。现在的问题是，在使用 316 等字段的$5 子字段前提下，905 字段的作用和存在意义是什么？

在现在的古籍机读目录格式当中，316 等字段的$5 子字段和 905 字段都是记录复本馆藏信息的。我们可以这样理解，905 字段是当时没有$5 子字段，为了记录馆藏信息而设置的。而现在的 316 等字段中的$5 子字段，是 IFLA 国际书目控制和国际 MARC 委员会（UBCIM）设置的国际通用的记录馆藏信息的子字段。即 905 字段和复本字段的$5 子字段的功能是重复的，需要研究解决。据了解，国家标准的《中国机读目录格式》（征求意见稿）已经不涉及 905 字段。

需要指出的是，当根据国家标准《中国机读目录格式》修订的新的古籍机读目录格式确定以后，已经投入使用的非 CNMARC 格式的古籍书目数据，均应尽早转换为 CNMARC 格式，以便纳入全国统一的古籍书目数据资源共享体系，发挥更大的作用和效益。

二、消除书目数据转换差错

制作好的古籍书目数据经过技术转换才能上网供读者检索。这个转换工作也是容易产生差错的环节，有时甚至使得原有书目数据面目全非。产生差错的原因主要有二：一是用于转换的专用程序主要是根据普通图书书目数据编制的，而古籍书目数据与普通图书书目数据在格式上有所不同，不加变动拿来直接使用容易产生新的问题。二是数据转换工作人员不熟悉古籍书目数据的细节变化和特征，有些他们认为没有问题的地方却容易出现差错。

以 2002 年 9 月 8 日国图网页上古籍书目数据为例分析归纳，主要问题如下：

1．不应该有的著录大项合并，如版本项与出版发行项，或者卡片显示中的出版发行项前没有大项符号。但个别有出版地的数据没有这个现象。

2．书目单和卡片显示不一致，如书目单上有多个著者，卡片显示就成了一个著者。书目单的两个著者之间有逗号，卡片显示中就没有了。

3．有些古籍是日本出版的，已经加了“日本”出版地，卡片显示又加了“CN”，这就错了。

4．上网古籍书目数据的卡片格式均没有册数。善本书目单的册数时有时无。

5．相关题名、著者不能实现自动连接检索，成了摆设，容易使读者产生误会。

6．业务注记也出现了问题，如：许多地方索书号有“部二”的，有些在书目单和卡片格式分别有“部一”、“部二”两条记录，有些则只有一条记录，极不统一。善本书目单的方括号中是空的，没有索书号和阅览地点。分馆古籍书目单方括号中的阅览地点时有时无，有许多书目数据的阅览地点为地方志家谱阅览室错成普通古籍阅览室。

我们在考虑消除书目数据转换差错时，应该重新权衡上网的古籍书目数据需要具备哪些著录项目，考虑上网古籍数据的技术处理与古籍机读格式如何取得一致。古籍书目数据经过技术处理上网以后，应该当即进行检查，及时发现、改正转换差错，并成为正常的网上古籍书目数据维护工作机制的首要组成部分。

三、审校著录项目

古籍书目数据经过著录、分类、校对、转换等环节上了网，由于种种原因（主要是古籍的复杂性），还会有各种内容差错存在。这就需要我们把定期审校网上古籍书目数据的著录项目、定期进行更新完善数据的工作作为古籍书目数据维护机制的重要组成部分。

我曾在网上看到这样一条书目数据：

潜川汪氏敬思流芳集：一卷，后集一卷/（明）汪静甫纂修.--抄本.--明

我直觉感到“静甫”更象是字号，不象是名字。经查，此书卷端题：“三十二世孙 永龄

静甫 续录”。书中正文“后集·三十二世”载：“永龄，字静甫，号东山，德洪公子，生大明弘治十四年辛酉十一月庚子初七日辛巳己亥时。子：文澜、文翰、文沛。”看来是编目员没有查到书中的作者小传，在“永龄静甫”四个字中选择作者名字时错了。

故此书责任说明应著录为：

（明）汪永龄续录

网上古籍书目数据的著录项目还会因为新老古籍著录规则的不同产生差错。如果书目数据是根据原书编制的，现在著录规则发生变化，应该重新审校有关的著录项目。如果书目数据是根据原编目卡片转录的，卡片的编目年代更早，现在的古籍著录规则与当年的规则相比肯定改变不小，更应该重新审校有关的著录项目，根据新的著录要求改补有关著录内容。如上例中的“纂修”改为“续录”，是因为原来著作方式的著录以“统一”为主，而现在著作方式的著录以“照录”为主。

审校网上古籍书目数据的著录项目，是一项较长时期的业务要求较高的工作项目，也是需要古籍书目数据库工作组织者认真考虑的重要问题。

四、调整古籍编码字段内容

调整古籍编码字段的内容是指中国古籍专用编码字段，如一般性数据等。从目前的古籍书目数据来看，古籍编码字段有三类：一是非 CNMARC 格式的古籍书目数据，其古籍编码字段各不相同。二是用 CNMARC 格式，但尚未用古籍机读目录格式，使用的是 105 字段（一般性数据），与普通图书一样，没有 140 和 141 字段。三是用古籍机读目录格式，使用的是 140 字段（一般性数据）和 141 字段（复本特征）。因为 140 和 141 字段不能很好地反映中国古籍的特点，目前在起草的国家标准《中国机读目录格式》拟用 193 和 194 自定义字段来替换 140 和 141 字段，两者的性质是一样的，内容有所不同。此国家标准发布后，上面所说的三类古籍编码字段都要统一调整到 193 和 194 字段上来。

193 字段（中国古籍——一般性数据）记录中国古籍文献原有的内容和物理形态特征的定长编码数据，主要元素内容有书籍代码（含内容类型、制作技术、墨色、刻工、出版者牌记、避讳、支撑材料等）和插图代码（书、全页图版、制作技术、墨色、支撑材料等）。194 字段（中国古籍——藏本形态特征）记录所藏的中国古籍文献现有的物理形态特征的定长编码数据，主要元素内容有装订材料代码、装帧形态代码、装订类型代码、合订代码、保护状况代码——封皮、保护状况代码——书籍主体。193 和 194 字段基本反映了中国古籍应记录定长编码数据的传统特点，也与 140 和 141 字段一样体现了“完整本著录原则”的基本内容，它把原有特征和现有特征区分开来，分别记录定长编码数据。

调整古籍编码字段内容是一项比较复杂的工作。记录 193 字段的内容基本上要依据原书重新制作。记录 194 字段的内容，如果记录中没有 141 字段，就需要依据原书重新制作，如果已经有 141 字段，只要明确各项元素之间的对应关系，可由程序自动控制完成转换工作。

五、完善文献连接字段

MARC 格式的文献连接字段是明确记录间相互关系、提高文献自动检索水平的技术手段。古籍机读目录格式的文献连接字段与此是一致的，因为中国古籍具有总体数量巨大、关联文献众多、内部层次复杂等特点，对于应用文献连接字段的需求是相当突出的。目前，之所以在古籍书目数据制作阶段文献连接字段方面存在一定的欠缺，主要原因是网络容量、人力和资金等方面的不足。古籍书目数据制作是有阶段性的，整改古籍书目数据则是一项长期的工作。我们应该根据网络容量、人力和资金条件的不断改善，在整改古籍书目数据工作中，注重完善文献连接字段。

根据个人的工作体会，笔者认为，完善文献连接字段主要应该注意下列几项内容：

1．制作丛书零种的总记录。建设古籍书目数据库时，因为条件所限没有制作丛书零种的总记录。当数据库中没有丛书零种的总记录时，丛书零种记录也不好做上连字段，连记录头标的层次级别代码也不好确定。所以，在整改古籍书目数据工作中，应根据积累的丛书零种档案，适时制作丛书零种的总记录，及时完善丛书零种记录的上连字段和层次级别代码。

2．补充古籍记录的相关连接字段。相关连接字段有两种：一是 470 字段（被评论作品）。本字段用于所编古籍向前与被它评论的古籍记录的连接，记录被评论的古籍文献的有关信息。200 字段记录含评论的古籍。若有多种被评论的古籍需要记录时，可将每种被评论的古籍分别记入一个 470 字段。二是 488 字段（其他相关作品）。本字段用于所编古籍与用其他 4--字段无法连接的另一个文献记录的连接。它连接的另一个文献和所编古籍的版本可以是相同的，也可以是不同的，包括古籍中所含有的具有检索意义的附刻、附录、书目、年谱等资料的书名等项目。补充古籍记录的相关连接字段，可以增加读者的相关古籍的检索种类和数量。

3．补齐连接字段中的记录标识号。当连接字段包含被连记录的记录标识号时，可以在连接字段中省略被连记录的其他字段内容。因为该记录标识号是唯一的，通过被连记录的记录标识号，可以找到被连记录的全部内容。由于条件所限，书目数据制作之初，连接字段没有包含被连记录的记录标识号，则被连记录的其他字段内容不能省略，而且，不能只写出被连记录的 200 字段（这样专指性不强），还应连上足以标志唯一被连记录的其他字段（尽可能少）内容。补齐连接字段中的记录标识号，可以提高读者的检索速度，精简记录字符数。

4．补做 856 字段（电子文件地址与检索）。本字段包含查找与本古籍记录有关的电子文件资源时所需的各项信息。这些信息包括电子文件有效的电子地址、电子文件名称及其查询电子文件的检索方式等。856 字段可用于查找与本记录揭示的古籍相关的电子版本或电子书影（用$u 子字段），具有古籍元数据的部分作用。在古籍的电子版本日益增加和古籍元数据尚未实际应用期间，856 字段的电子文献检索意义是显而易见的，但它的制作是比较花工夫的。

六、补齐系统外字

录入古籍书目数据时，现有字库里没有的字是经常遇到的。为了解决这个问题，古籍机读目录格式设置了 393 字段（系统外字附注）。本字段包含记录录入时字符集所缺字符（在出现该字符的位置已用“〓”表示）的结构和读音描述。但这种补救方式毕竟是权宜之计，既有碍书目数据的完整性，如果出现在检索字段（是常有的），还会造成难以检索的麻烦。所以在古籍书目数据整改阶段，应该使用 ISO 10646 Level 3（UNICODE）国际汉字大字符集（它应建立新字补入机制），补齐原来的系统外字（含清除“〓”和 393 字段），提高读者对古籍文献的查全率。

七、贯彻完整本著录原则

《中国机读目录格式使用手册》（2001 年修订版）在“141 编码数据字段：古籍——复本特征”中指出：古籍复本在此处泛指古籍的原件、原版、原稿、副本、复制品、抄件、印件等，强调的是“一件”。原稿、抄件等也应视为复本，这就与传统的复本含义不同了。传统的复本概念专指同版书第二部及其以后各部书，不包括稿本、抄本等和同版书的第一部。复本的内涵为什么会有这种变化？关于这个问题，IFLA 国际书目控制和国际 MARC 委员会（UBCIM）主持编写的《UNIMARC 指南 3．古籍（善本）》指出： “理想本或完整本和残缺本”作为 UNIMARC 模式的著录依据需要明确区分，以便适合日益突出的古籍机读目录和网上联机编目的需要。《UNIMARC 指南 3．古籍（善本）》认为：

1．古籍的“理想本或完整本”就是进入发行或流传领域以前的图书，它所具有的是古籍的原始版本特征，还没有“打上”发行或流传领域的烙印。

2．编目员手头所编的古籍都是复本。复本包含了古籍流传、收藏过程中产生新的特征和变化，也包括残缺本。也可以这样理解，一部手稿，在作者手中是“理想本或完整本”，到编目员手中就是“复本”，按“复本”的著录要求处理，不管它是否“打上”发行或流传领域的烙印。

3．“理想本或完整本”的原始版本特征应著录在基本著录字段，复本情况（含索书号）

全部著录在 141、316、317 和 318 等复本情况著录字段。

4.“理想本或完整本”的基本著录字段内容适合所有复本或反映所有复本的原始情况，每种基本情况只有一个字段（指 200、205、210、215 和 140 等字段），不能重复，个别字段可以空缺。复本字段可以视复本数量任意添加，但同一复本的有关复本字段必须对应。

此间，“理想本或完整本”就是我们所熟悉的古籍制作、发行之初的原始本（全本）。

《UNIMARC 指南》之所以如此规定，主要是基于下面的考虑：

当古籍的书目著录局限于本文献类型之内（未与普通图书的著录统一起来），局限于本馆范围之内时，手头所编古籍的各方面版本特征暂可以不加区分地成为 UNIMARC 模式的基本著录依据，同一原始版本的各个复本由于后续特征不同而分别做一条数据记录。当进入各种文献类型（包括普通图书和古籍等）的著录规则需要统一、网上各馆古籍联机合作编目的时代前提下，手头所编古籍的“各方面”版本特征就需要明确区分为原始版本特征和后续版本特征，分别进行著录，以便达到既集中（同一版本）又有区别（不同复本）的目的。这一点，在我们传统的联合目录中早已经反映出来了。

IFLA 国际书目控制和国际 MARC 委员会（UBCIM）认为：在网上联机编目中贯彻这个原则，可以建立“在完整本著录前提下，所有入编古籍一律平等（都是复本），同一版本的不同复本互相关系清晰”的格局。

在目前古籍书目数据当中，由于后续版本特征的不同有关复本另做记录的原因，除了当时机读目录格式没有提供集中全部复本的解决办法以外，还因为考虑古籍的后续版本特征在

书目史上占有比较重要的地位。下面的古籍书目数据引自 2002 年 9 月 8 日的国家图书馆网页，它反映了由于后续版本特征的不同有关复本另做记录的基本情况：

1．钤印等特征不同：

书 名	版 本	记 录1	记 录2
初学记	明万历二十五至二十六年维扬陈大科刻本	8册，钤“积余秘籍识者宝之”印	12册，钤“会稽李氏困学楼藏书印”、“湖唐林馆山民”等印
初学记	明嘉靖二十三年凤阳朱胤栘翻刻本	12册，钤“柯逢时印”、 “潘炳辉印”等印	12册，钤“吴兴姚伯号觐元鉴藏书画图籍之印”

2．册数不同：

书 名	版 本	记录1	记录2	记录3	记录4	记录5	记录6	记录7
初学记	明嘉靖十年安国桂坡馆刻本	12册	6册	12册	1册	5册		
史记	明崇祯十四年毛氏汲古阁刻本	16册	8册	12册	10册	10册	14册	8册

3．全缺不同：

书 名	版 本	记录1	记 录2	记 录3
钦定四库全书总目	清乾隆武英殿木活字本	120册	112册，缺卷首1-4卷、正文卷1-12	1册，存卷首第2卷

上面所摘引的同一版本的不同复本另做记录现象是比较普遍的，它既与普通图书的著录原则不统一，也不符合 IFLA 国际书目控制和国际 MARC 委员会（UBCIM）提出的有关“完整本著录”的最新解决方案，也是不适合以后国际古籍书目数据方面的交流与合作。

为了在古籍书目数据中准确、合理地揭示同一版本的所有复本特征（如：批校题跋、藏章、挖补、人为合订、残缺、修补、换皮订线、消毒等），适应古籍网上联机编目的大环境需要，与 UNIMARC 的原则要求接轨，我们应该认真审视由于“完整本著录原则”产生的有关问题，供进行古籍书目数据整改工作时参考：

1．古籍的已有书目数据是否应根据 UNIMARC 的“完整本著录原则”进行整改？

回答应该是肯定的。我们不能强调古籍的特殊价值，而与普通图书的著录规则不统一，违背 UNIMARC 的“完整本著录原则”，最后导致在国际上进行古籍书目数据交流和合作方面产生困难和障碍。古籍的特殊价值，应该通过内容特征和形式特征的全面著录加以反映与体现，而不是采用增加记录数量的办法。在内部馆藏书目记录中需要增加复本记录时，可以通过特定程序，在统一记录中抽取相关字段组合成不同的复本记录。

2．古籍完整本和复本特征应该分别记录在哪些字段？

古籍完整本的特征著录在 140 或 193、200、205、215、305、306、307 等字段。没有完

整本时（如没有全本），参照完整本的书目著录进行手头古籍复本的编目。

复本的全部特点包括残缺情况必须在 141 或 194、316、317 和 318 等字段中记录。每部复本必须著录 141 或 194 和 316 等复本字段。所有复本字段用$5 子字段记录收藏该复本的机构代码和索书号。如找不到完整本的书目著录或考证不出完整本的有关特征，200 字段的$e 其他题名信息、$f 第一责任说明和$g 其他责任说明以及 215 字段可以空缺，但其他项目必须按 ISBD（A）(《国际标准书目著录（古籍）》) 的规定进行著录，如从手头所编古籍以外考证得来的信息，著录时应加“[ ]”。

3．如果复本特征附注记入 316 等字段，当复本特征较多时，316 等字段应该任何使用？

每部复本原则上只能使用一次 141 或 194、316、317 和 318 等字段。当某部复本的后续特征较多时，316 字段可以重复，分别按册数、存缺、批校、钤印、其他后续特征的顺序进行著录，但其$5 子字段（索书号必备）必须一致，以免引起混乱。

八、统一组建古籍书目数据库

统一组建古籍书目数据库，目前暂就国图的情形而言。国图的网上古籍书目数据目前已达 18 万条，分为“普通古籍库”、“善本古籍库”和“方志家谱库”。这主要是依据藏书部门划分的，与读者的古籍利用需求不尽一致，检索不太便捷。“ 统一组建”的含义有二：一是目前普通古籍和善本古籍的书目数据格式因是不同部门制作的，略有差异，应该统一。二是古籍书目数据库分合应该统一。普通古籍和善本古籍都是古籍，两库所藏古籍有交叉，不少同一版本的不同复本又分藏两库（如上例的《初学记》桂坡馆本），应该合为一库。目前的善本古籍库中含有方志家谱，应该分出来，并入方志家谱库，便于读者使用。以前因为普通古籍和善本古籍分属不同部门，手工目录不在一起，读者利用不便。现在已经建立网上古籍书目数据库，电子目录就没有理由再分割了。

以上说明的八项古籍书目数据整改工作内容，第二项（消除书目数据转换差错）应该马上进行，第三（审校著录项目）、第五（完善文献连接字段）、第六（补齐系统外字）和第八项（统一组建古籍书目数据库）只要条件（书目数据制作告一段落、人力、连接软件和字库）具备即可进行，第一（修订古籍机读目录格式）、第四（调整古籍编码字段内容）和第七项（贯彻完整本著录原则）要等有关规则，如国家标准《中国机读目录格式》、新的《中国文献编目规则》和《汉语文古籍机读目录格式》确定以后再进行，但我们现在就应该开始这些整改工作的分析、研究。

数据维护工作是保证上网数据正常运行的必要手段。数据整改工作则是数据维护的首要工作内容。鉴于古籍书目数据的复杂性和重要性，我们不但要抓紧制作工作，更应该注重整改工作，因为制作工作仅仅解决古籍书目数据有没有的问题，而整改工作则关系到古籍书目数据能否发挥应有作用的问题。