中华书局的古籍数字化之路综述

中华文化基因工程服务联盟
移动媒体与文化计算北京市重点实验室

一、中华经典古籍库

（一）理念

在古籍数字化探索中，中华书局形成了一系列的理念：

首先，古籍数字化必须以古籍整理为基础。目前的传世古籍存在大量问题，如果不整理，难以投入广泛使用。在整理古籍时，必须按照规范和标准来整理，这样才能保证质量，才能整理出被广泛引用的善本。但是当前的一些古籍数据库无法被直接征引，低于学术应有的水平，也不符合学者的期待。

其次，古籍数据必须优质合法。古籍的现代整理已有百年历史，创造了丰富的可供数字化的成果。初步统计，民国整理约 1 万余种，新中国整理出版古近 3 万种，其中可以数字化的比较好的版本约 3000 种，许多是富有盛誉的古籍整理精品。同时古籍整理作品还拥有著作权，被中国的著作权法和司法实践确认，既包括纸质著作权也包括信息网络传播权。

再次，古籍数字化的服务必须符合用户需求和学术规范。数据库的服务功能虽然是技术创新，而且也依靠技术创新来提升古籍整理的水平和使用效率，但也一定要符合古籍使用者的需要和学术规范。

最后，古籍数字化的技术标准必须符合中文古籍特点。古籍数字化流程中，数据采集标准规范、图像采集处理规范、文字采集规范、资源内容表示规范、古籍用字标准、专有名词标准、整理本古籍元数据标准、整理本古籍文献数据标注标准等等，都必须符合中文古籍的特点。

（二）起源

2014 年 6 月 10 日，“中华经典古籍库”是中华书局进行古籍的首度数字化，收录包括“二十四史”及《清史稿》《资治通鉴》等经典系列在内的权威整理本，一期收录 300 种，计 2 亿余字。产品不仅提供了保留全部整理成果的数字文本，更实现了文本与原书图像的一一对照，并能自动生成引用格式，除支持全文检索外，还添加了独具特色的人名异称关联检索等。

2003 年，中华书局成立了“古籍资源开发部”，负责建设“中华古籍语料库”项

目，正式开启了古籍数字化的工作。古籍资源开发部最主要的工作是数字化编辑加工，即将中华书局铅排古籍整理图书通过 OCR 采集等手段数字化，并按照语料库的要求编辑加工成 XML 格式的文件。通过几年建设，完成了 3 亿字整理本古籍的数字化工作，籍此书局建立了一整套数字加工标准和流程管理规范。古籍资源部在数据加工的同时，也开发了一套系统用来编辑、标引、存储、检索、统计古籍数据。

在整理本古籍的数字化工作中，最个性化的困难是计算机用字的处理。“中华经典古籍库”在系统里使用了 Unicode 编码字符集，从基本集到扩展 E 集，共有 8 万余个国际标准编码的汉字。到目前为止，额外造了 3.1 万个字符集以外的字，使这些字具有检索功能，并实现灵活的样式显示，在 PC 端和移动端可提供更好的阅读和检索体验。

（三）产品化

2014—2015 年，“中华经典古籍库”的局域网版是数字化的主要产品和收入来源，这和产品定位在国内机构用户是紧密相关的。大多机构希望一次性付费使用，并买到实体，而不是按年付费订阅。

但是局域网版具有很多劣势：一是海外用户基本不会采购，他们不愿安装软件到本地，更多的是希望通过浏览器在线访问；二是数据库的采购通常伴随大量的试用过程，局域网版需要给用户上门安装，如果不采购还要再撤回，试用效率低而且成本高。

2015 年底，发布了在线版，可以通过网络授权访问。短短一年里，在线版已经在 100 多个机构开通试用，北美地区的哈佛、耶鲁、普林斯顿、哥伦比亚等几所大学都购买了在线产品。在线版的试用不仅让更多机构了解到了“中华经典古籍库”，而且对局域网版的销售还起到了促进作用。

2016 年 4 月 23 日，在中华书局读者开放日上，发布微信版“中华经典古籍库”。这是在社交移动平台第一次出现大规模的古籍资源，读者可以随时随地的阅读检索，分享内容。

在产品里还为用户提供了一些必要工具，比如联机字典、历史纪年换算、关联字表查询等等；利用工具书和原书后的索引，制作了人名异称的关联检索，当用户检索时输入一个人名，系统会提示这个人物在文献中的其他称谓，

比如曹操，系统会提示孟德、魏武帝、阿瞞、吉利等等，便于用户提高检索的查全率。

（四）用户推广

“中华经典古籍库”无论是局域网版还是在线版，都是面向机构用户销售的产品。用户构成主要有以下几类：高校图书馆及专业院系、公共图书馆、党政机关、出版社、研究机构及博物馆、其他民间机构、海外机构（主要是大学和一些国家图书馆）。根据调研，古籍库机构版的潜在用户不下千家。但是经过两年的推广，古籍库的试用用户只有上百家，这种一对一的面向机构推广的模式效率不高。有90%以上的读者，无缘接触到“中华经典古籍库”，甚至不知道中华书局有了古籍数字化成果。另一方面，面向机构的产品存在一个天然的问题，采购者和使用者往往是两个群体，因此不太容易接触到真正的用户，用户也无法顺畅地表达对产品的反馈。为了解决这些问题，需要一个面向个人的产品，还要具备高效的传播方式，用户能够很方便对产品沟通，自然就选中了微信作为数据库的载体。

微信产品让中华书局第一次真正地面向读者：通过后台的统计分析功能，可以了解到用户检索和阅读哪些内容、什么时间使用数据库、哪些地方的用户最多、他们操作方式是什么等等。很多读者本着对中华书局及其产品的信任，在注册时提供了完整的注册信息，包括专业、职业、联系方式等等，这让中华书局有了更加具体的用户画像。通过微信的二维码关注功能，在不同活动、不同推广媒体和场合投放的二维码，可以明确区分出用户群体和传播渠道。微信用户数据对于中华书局明确产品的发展方向和提升营销的针对性都起到了关键作用。真正定位到“人”，是微信产品最核心的价值。认识到这一点后，中华书局与高校用户开展合作，将微信版账号赠送给在校学生使用。对于学生来说，他们获取到了一大批免费优质资源；对于学校来说，已经购买的机构版新增了额外的增值服务；对于中华书局来说，得到了一大批潜在用户。

（五）平台化

中华书局在发展了两年产品后，深切感到要想真正做到融合发展，推动出版的转型升级，靠现有的产品线是不够的。很多根本性的问题没有解决：

第一、古籍整理作品通过整理者和编辑的努力，内容质量远高于社会其他古籍资源，但是在数量上具有天然劣势，无法满足用户在更大范围内检索文献的需求。

第二、现有数字产品的模式实际上是纸书的附属物，从内容到版权，都受制于纸质图书的出版，还远达不到产业转型升级的要求。

第三、在互联网时代，很多与内容相关的产品并不是依赖对现有内容的数字化来完成的，而是通过用户自己产生内容，迅速扩张，维基百科、知乎等都是很典型的例子。

出于这些原因，感到发展产品只是中华书局工作一部分，更重要的是通过互联网搭建一个平台，提供一个古籍整理的新模式，加速古籍的整理速度。

“中华古籍整理出版资源平台”力图打通数字与出版的双向通路：古籍整理出版物可以通过数字产品的形式发布，整理平台也可以通过数字化产生整理作品，可直接在线发布，还可以提供给出版社纸质出版。平台提供了古籍从整理到发布的一系列流程：

（1）该平台提供了一个古籍书目系统，包括了从版刻书到整理本一系列的古籍目录，可以让用户方便的检索古籍书目信息，并且了解整理出版情况。不仅能达到检索古籍书目的目的，还能够依照中国古籍的整理情况，进行古籍整理的规划工作。

（2）提供了版刻书调阅系统，涵盖大量的版刻图书资源，以原版扫描的形式提供，作为用户整理古籍的底本和校本使用，也可作为其他的整理参考。

（3）作为一个古籍整理平台，提供了自动校勘和辅助标点功能，利用后台的数据支持，为用户整理古籍提供大量的参考资料和已有整理成果。

（4）工具书与知识单元查询系统，深度嵌入到整理平台，为整理者提供必要的知识提示。

（5）成果发布系统，可以将在线的整理成果直接发布，供读者使用。同时，平台还具备一个约稿系统，可以发布需要整理古籍的信息，采用众包的形式，由读者共同整理完成。该平台的设计从根本上要实现古籍的在线整理和发布，通过众包与多人协作，提高古籍整理的速度。

在平台下面，学术期刊库、碑刻墓志库、小学文献库等等多个专业子库也在研发中，既可以为古籍整理者提供资料支持，也可以作为单独产品运营。中华书局希望通过平台的建设，将进一步打通读者和作者之间的关系，通过互联与协作，推动古籍整理事业的发展。

（六）成功经验

中华文化数字化工程体系的整体构建需要国家层面的顶层设计和布局，无论是要实施的项目还是制定的细则，都要有整体系统的规划。有了数字化这个前提，再加上一些技术分析手段，就为研究中华民族文化奠定了基础，在这个基础上可以更好地制定我国文化发展的整体战略，不仅包括文化产业的发展，还包括文化的建设和发展。唯其如此，才能明确要往我们的文化产品里植入哪些文化元素，对外传播什么文化内容。

在做数字化过程中发现，资料存储是一个大的问题。这么多数字化成果，怎样建立一个数据库存下来，并且能够查询、管理，更重要的是应用，这是下一步要做的。只有把文化资源数字化，把这些碎片标签化变成素材库，才能在更多领域应用，但是目前仅仅做到数字化、做成数据库提供给一些学者和其他很小范围的使用者使用。如果大范围的应用，把这些数字资源创造、加工、生产，然后去传播、消费，还有很长的路要走。这个过程面临资金问题、知识产权问题、标准问题、应用渠道狭窄等诸多制约。

技术在数字出版领域占据重要作用。这里所说的技术并不是单指计算机技术，还包括提供数字服务的一切相关技术、标准。以中华书局的产品为例，包括了数字化的相关标准、超过 10 万字的古籍字表和属性数据库、不断完善的汉字关联表、准确的历史纪年换算工具、几十万的专名词表、在线显示超大字符集和版权保护技术等等，这些都是在数字化过程中不断发展出来的。2016 年底，古联公司组建了“古籍数字化与知识工程重点实验”，成为首批新闻出版业科技与标准重点实验室，在“古籍数字化汉字处理”“古籍文本自然语言处理与语义关联”“古籍知识组织体系建设”“古籍资源知识库构建”“古籍整理自动化”五个角度进行深入研究，这些将来都是构成古籍整理数字化工作最核心的技术。在资源量达到一定级别的时候，技术的价值将越来越充分地体现出来。

（七）近期成果

“中华经典古籍库”持续更新数据，每年推出一辑数据包，持续收录新出版的优秀整理本古籍，在保证质量的基础上有序扩充数据量，同时不断进行数据的修订与完善。此外，人名异称关联表与联机字典也在不断扩充中，将会为用户提供更丰富的服务。中华书局也将不断扩展产品线，通过开发更多的专题库、小型库以满足不同用户的需求。商周铜器铭文知识库将在前期资源整理的基础上逐步实现产品化，而中华书局的第二个大型数据库产品《中华基本史籍知识库》也已经启动，该产品将在古籍库的基础上，收入学术著作及工具书，借鉴“史籍分析系统”项目的建设经验，建立人物、时间、地点等史籍知识元间的关联，可视化地展示其语义关系，为学者提供更为专业的知识服务。“中华书局最终目标是系统地完成整理本古籍的数字化，搭建自成体系的知识网络，让知识之间建立链接，打破专家与读者之间的知识和信息壁垒，为社会提供真正有价值的古籍数字化产品”。

中华书局希望，“中华经典古籍库”的出版能促进古籍数据库市场完善。中华书局会坚持“专业、优质”的出版理念，稳扎稳打，不盲目跟风，专注于提供最佳的数字服务，为中国优秀的传统文化的有效传播做出实实在在的贡献。

二、未来趋势

（一）科技化

在数字化的国家战略中，已经包括文化数字化。脑神经网络计划、3D 打印机、4D 打印机等项目未来要成为数字化的主力，称为“制造创新国家网络”。未来国家综合实力的竞争将不仅在军事、经济等传统层面，还将体现在信息、文化等领域。当前我国对文化要素的数字化转换技术仍相对落后，大家应抓住数字化浪潮的机会，在未来的国际竞争中掌握主动权。

（二）体系化

文化资源数字化工程确实面临一些问题，一是由谁来转化，二是如何转化。文化资源大多属于文化单位，多数承担的是管理职责，深度研究和技术处理这方面还面临着很多问题，比如资金来源、资源共享等。其中大量是科技之外的问题、工程之外的问题，既与现行的联动机制，包括成果的分享、利益的共享有关，也有知识产权的问题，更有深层次的体制机制问题。目前，文化资源数字信息大量

分散形成一个个信息孤岛，地方没有一个统一的标准，统一整合对接存在一些问题。未来，要有目标性，形成一个主线、一个体系。

（三）标准化

现在方方面面都在做文化资源的数字化工程，包括博物馆、图书馆、非遗等形式多样的数字化，但是很多数字化以后的资源也是在各自单位里分别存放，很难放在一个库里。当前，更重要的是把这些资源有效整合利用起来，建立一个共享共建机制，使得数字化资源通过某种技术平台更好地整合起来。这需要制定一个规划，明确用什么样的标准做数字化，这是一个模式，也是一个机制，从而建立技术上的共享公共平台。

（四）**管理机制优化**

解决文化数字化中的诸多问题，要从如何破解现在落后的管理机制上入手。比如资金投入问题，国家是否要给资金支持？国家要给支持，给政策，也要监督。应该划小核算单位，把物权要管清、管好，实现所有数字资源的产权化。再比如数据采集，要采用市场化运作模式，要具体分析项目到底该怎么做。国家可以用启动资金来实现政策导向，主要还是要靠政策、靠监管把这个事情做起来。

三.更广范围应用

把文化数字化工程提升到国家战略的层面，就是为中华文化在信息化时代创造一个新的文化再生产体系，让文化资源以适合的数字化形式呈现出来，重新引起人们的关注。

在这个过程中，文化资源的数字化转化环节无疑是最难的。比如文物，中国所有的青铜器都在博物馆仓库里，没有数据体系，将这些数量巨大的青铜器信息进行数字化转换困难极大。

另一个难点在于，现在的整体文化研究系统中，核心价值领域系统和数字化仍有很远的距离。现在仍处于“写文化”的时代，学术界的大部分研究成果呈现在纸面上才能被认可。将教育系统的整个知识体系进行数字化改造，是文化数字化工程的第一步。

在中华文化数字化工程整个进程中，一个最大的问题是基本没有文化数字化的规划。现在，我国各地文化资源非常分散，非常破碎。在发展中，首先要有规划，要盘点中华文化的库存，盘点成什么样子要有标准，根据这个标准要

分类，从而固定一些产品属于哪一类。根据划分的种类，要有知识产权保护的整体意识，还要制定知识产权保护规划。