地方志数字化加工规范及研制情况介绍

张毅

据统计，现存旧志（1949 年以前编撰或出版，含专志）超过 1.5 万种、3 万个版本，约 1 千万个筒子页。目前，已有三分之二完成图像采集，而文本化不超过三分之一。作为重要的学术资源，地方志数字化的数量与质量都不能满足学界要求。但要解决上述问题，实现地方志大规模数字化，就必须建立相对完善的标准规范体系，形成高效的地方志数字化工艺流程。

国内外许多图书馆、高校、研究所等文献典藏机构都建设了文献数字化项目并取得了一定的成绩，许多项目在发布资源、便利用户使用的同时，还发布了其项目建设标准规范。国内具有代表性的项目如：“中国数字图书馆标准与规范建设”项目（CDLS）、“国家数字图书馆”标准规范（中国国家图书馆）、“中国高等数字图书馆（CADLIS）”技术标准与规范、“北京大学数字图书馆”标准规范、“大学数字图书馆国际合作计划”标准规范、（台湾）“数位典藏和数位学习”项目标准规范等。地方志数字化可参考借鉴者目前只有地方志元数据、图像数据的标准规范，而没有地方志索引、文本、外字处理、资源库等标准规范。在已有或在建的地方志数字化项目中，大量使用数字化厂商或文献收藏机构自定义的数据标准。这就使得地方志数字资源质量参差不齐，数据格式和加工方法各不相同，给数据整合与资源共享带来了巨大的困难。

项目组结合长期以来地方志数字化的工作经验，依据地方志文献特性和项目需求设计并研制了 5 类 8 个地方志数字化加工规范，力图构建一套相对完整的、行之有效的地方志数字化标准规范体系。包括文字处理规范 2 个：汉字集外字描述规范和文字认同描述规范；元数据规范 2 个：地方志专门元数据规范和地方志卷目数据标引规范；对象数据规范 2 个：地方志图像数据规范和地方志文本数据规范；语料数据规范 1 个，地方志语料数据规范；知识库数据规范 1 个，古今地名数据规范。一方面，对现行的规范进行修订，使其更具实用性。如地方志元数据规范。以往的元数据规范多侧重于书目数据，对数字化对象的描述非常详尽，但缺少数字化加工过程中管理信息的著录，如分辨率、压缩率、加工方式（扫描、拍照、缩微转换）等，而这些信息对于数字资源又是非常重要的。另一方面，填补空白，建立新规范，

如汉字集外字描述规范、文字认同描述规范等，着力解决数字化项目（厂商）在处理相应问题时各行其是的问题。

以下对 8 个地方志数字化加工规范进行简要介绍。地方志元数据规范用于整体描述地方志图像数据和文本数据，内容包括书目信息、数字化参数、版权信息、用户使用提示等。地方志卷目数据标引规范用于描述地方志图像数据的卷次结构，内容包括卷目信息、层次结构、分类、起始页码等。地方志图像数据规范用于描述地方志图像及规范图像采集方法，内容包括图像类型、画幅内必备要素、图像参数、数字化设备参数、图像压缩、色彩管理等。汉字集外字描述规范用于描述地方志文本数据中的集外字，内容包括集外字定位、集外字图像（或图形）、IDS 描述等。文字认同描述规范用于描述地方志文本数据中的文字认同情况，内容包括认同字定位、认同方式、认同类型等。地方志文本数据规范描述地方志文本数据的基本结构和全文文本描述方式，内容包括文本数据格式、文字描述、版式描述、图形图像描述等。地方志语料数据规范描述地方志语料数据的基本结构和标引方式，内容包括文本、文字描述、版式描述、图形图像描述等。中国古今地名数据描述规范规定了中国古今地名数据的内容、构成要素及各要素的描述规则，适用于中国古今地名数据库的建立及格式交换。

标准规范的参研人员众多，多是参与地方志数字化工作的直接参与者，主要来自国家图书馆、汉王科技股份有限公司和中国传媒大学。早在 2013 年项目申报阶段，标准规范的设计、准备工作既已开始。项目正式立项后便着手规范的研制工作，每个规范采用专人负责、团队不定期讨论的方式，辅以实时反馈机制，保证了 8 个规范有条不紊的进行。在一些重要节点，如初稿完成、草稿拟定后，组织、召开专家咨询会，听取专家的意见和建议。同时，结合项目进度，将已完成的标准规范推至应用端，测试其适用性。

至 2017 年底，8 个标准最终定稿，并完成了应用指南的编纂。目前，8 个地方志数字化加工规范和应用指南作为本项目研究成果已由学苑出版社结集出版。此外，8 个标准规范作为汉王科技股份有限公司的企业标准，在企业标准信息公共服务平台 http://www.cpbz.gov.cn/ 发布。

依据地方志数字化加工规范，项目组还设计完成了图像自动拼接软件、集外字描述软件、语料数据加工软件和古今地名数据加工软件。

在 8 个地方志数字化加工规范中，项目组经过反复讨论，决定选取其中 4 个申报文化行业标准规范，包括中国古今地名数据描述规范、文字认同描述规范、汉字集外字描述规范和地方志文本数据规范。在文化行业标准草案申报过程中，文化部全国图书馆标准化技术委员会的专家认为“地方志文本数据规范”的适用范围较小，应将适用范围扩展到汉文古籍。经

项目组反复研究，认为地方志具有汉文古籍的普遍特征，于是对“地方志文本数据规范”进行了必要的修改，最终申报“汉文古籍文本数据规范”。项目组 2016 年申报了汉语文古籍文字认同描述规范（WH2016-01）和中国古今地名数据描述规范（WH2016-05）；2017 年申报了汉文古籍集外字描述规范（WH2017-03）和汉文古籍文本数据规范（WH2017-04）。目前，“中国古今地名数据描述规范”已在网上公示，“汉语文古籍文字认同描述规范”在进行专家审核，“汉文古籍文本数据规范”和“汉文古籍集外字描述规范”在进行草案修订。