国家数字图书馆工程标准规范成果

国家图书馆文本数据加工标准和操作指南

龙　伟　罗云川　主编

6-65
3

九江学院图书馆
1554611
1613268

国家图书馆文本数据加工标准和操作指南

龙　伟　罗云川　主编

G250.76-65
18763
九江学院图书馆
藏书章
不外借

國家圖書館出版社

图书在版编目(CIP)数据

国家图书馆文本数据加工标准和操作指南/龙伟,罗云川主编. —北京:国家图书馆出版社,2012.8

(国家数字图书馆工程标准规范成果)

ISBN 978-7-5013-4819-0

Ⅰ.①国… Ⅱ.①龙… ②罗… Ⅲ.①中国国家图书馆—文本—数据处理—规范 Ⅳ.①G255-65

中国版本图书馆 CIP 数据核字(2012)第 158177 号

责任编辑:高爽

书名 国家图书馆文本数据加工标准和操作指南

著者 龙伟 罗云川 主编

出版 国家图书馆出版社(原北京图书馆出版社)
(100034 北京市西城区文津街 7 号)

发行 010-66139745 66151313 66175620 66126153
66174391(传真) 66126156(门市部)

E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com→投稿中心

经销 新华书店

印刷 北京佳顺印务有限公司

开本 880×1230(毫米) 1/32

印张 4.25

版次 2012 年 8 月第 1 版 2012 年 8 月第 1 次印刷

字数 90 千字

书号 ISBN 978-7-5013-4819-0

定价 35.00 元

丛书编委会

主　编：国家图书馆

编委会：

主　任：周和平

执行副主任：詹福瑞

副主任：陈　力　魏大威

成　员（按姓氏拼音排名）：卜书庆　贺　燕　蒋宇弘

梁蕙玮　龙　伟　吕淑萍　申晓娟　苏品红

汪东波　王文玲　王　洋　杨东波　翟喜奎

赵　悦　周　晨

本书编委会

主　编：龙　伟　罗云川

编　写：赵四友　肖　禹　李鹏云　韩　超　王　浩
　　蒋卫东　陈艳平　袁海滨

此为试读，需要完整PDF请访问：www.ertongbook.com

总　序

数字图书馆涵盖多个分布式、超大规模、可互操作的异构多媒体资源库群，面向社会公众提供全方位的知识服务。它既是知识网络，又是知识中心，同时也是一套完整的知识定位系统，并将成为未来社会公共信息的中心和枢纽。数字图书馆建设的最终目标是实现对人类知识的普遍存取，使任何群体、任何个人都能与人类知识宝库近在咫尺，随时随地从中受益，从而最终消除人们在信息获取方面的不平等。“国家图书馆二期工程暨国家数字图书馆工程”是国家“十五”重点文化建设项目，由国家图书馆主持建设，其中国家数字图书馆工程的建设内容主要包括硬件基础平台、数字图书馆应用系统和数字图书馆标准规范体系。

标准规范作为数字图书馆建设的基础，是开发利用与共建共享资源的基本保障，是保证数字图书馆的资源和服务在整个数字信息环境中可利用、可互操作和可持续发展的基础。因此，在数字图书馆建设中，应坚持标准规范建设先行的原则。国家数字图书馆标准规范体系建设围绕数字资源生命周期为主线进行构建，涉及数字图书馆建设过程中所需要的主要标准，涵盖数字内容创建、数字对象描述、数字资源组织管理、数字资源服务、数字资源长期保存五个环节，共计三十余项标准。

在国家数字图书馆标准规范建设中，国家图书馆本着合作、开放、共建的原则，引入有相关标准研制及实施经验的文献信息机构、科研机构以及企业单位承担标准规范的研制工作，这就使得国家数字图书馆标准规范的研制能够充分依托国家图书馆及各研制单位数字图书馆建设的实践与研究，使国家数字图书馆的标准规范成果具有广泛的开放性与适用性。本次出版的系列成果均经过国家图书馆验收、网上公开质询以及业界专家验收等多个验收环节，确保了标准规范成果的科学性及实用性。

目前，国内数字图书馆标准规范尚处于研究与探索性应用阶段，国家图书馆担负的职责与任务决定了我们在数字图书馆标准规范建设方面具有的责任。此次将国家数字图书馆工程标准规范研制成果付梓出版，将为其他图书馆、数字图书馆建设及相关行业数字资源建设与服务提供建设规范依据，对于推广国家数字图书馆建设成果，提高我国数字图书馆建设标准化水平，促进数字资源与服务的共建共享具有重要意义。

国家图书馆馆长　周和平

2010 年 8 月

前 言

文本数据是数字图书馆资源建设中最主要、最基础的资源类型。所谓文本是指通过文字、符号形式所表现、传递的信息。据统计，在海量信息资源当中，有70%多的信息资源的表现形式是文本方式，有80%多的信息内容的获取来自于文本数据，可见，文本数据是人们信息传递和交互的最主要方式。图书馆收藏的文献中大部分是文本形式，如图书、期刊、报纸等。随着我国信息化水平的发展，人们对文本信息阅读的功能性和精确性的要求不断提高，图书馆文本数据加工工作的规范化要求更加具有紧迫性。

近年来，国内大部分图书馆都开展了文本数据加工工作，但由于各机构选择的标准不同，甚至同一机构的不同文本数据加工项目所依据的加工标准也不一致，文本数据的加工技术和保存方式的差异性很大。国外图书馆及其他信息保存机构同样面临着大量的文本数据加工任务，如美国国会图书馆 American Memory，澳大利亚国家图书馆藏品数字化，哈佛大学图书馆数字化项目。很多大型机构一般将制订标准规范工作置于最先阶段，目前正在跨机构进行合作的美国 FADI（Federal Agencies Digitization Guidelines Initiative）制定出事实行业标准，用以指导合作项目。为保持与国际先进数字图书馆发展水平的一致，国家图书馆密切关注世界各国数字图书馆项目中标准规范的应用

与发展;关注国际标准化组织 ISO,尤其是 TC46(Information and documentation)、TC171(Document management applications)等技术委员会,以及美国国家信息标准化委员会等主要国家标准化组织在数字图书馆相关标准方面的发展状况与趋势,在跟踪研究的基础上,制订适合我国数字图书馆数字资源建设的标准规范。

2008 年,国家数字图书馆工程标准规范项目采取竞争性谈判方式向社会各界发出参与研制的邀请。文化部全国文化信息资源建设管理中心受国家图书馆委托,承担了国家数字图书馆标准规范项目"文本数据加工标准与工作规范"的研制。国家图书馆"文本数据加工标准与工作规范"项目,是国家数字图书馆工程标准规范建设项目之一,项目的主要研究任务是根据国家图书馆文本文献和数字资源的情况,参考数字加工领域的最新技术成果,研制适合于国家图书馆的文本型数据的加工标准及其指导标准的操作指南。

文化部全国文化信息资源建设管理中心作为全国文化信息资源共享工程的国家中心,对文本、视频等数字资源具有多年的实践加工和服务经验,为顺利完成研制任务,与国家图书馆成立了专门的项目研制小组。项目研制小组对国内外主要的数字图书馆资源建设项目和国际通用的数字资源编码和文件格式标准进行了较全面的调研,根据国家图书馆资源建设现状,分析了各类建设项目和文本数字资源,在此基础上完成研制国家图书馆文本数据加工标准与工作规范及其操作指南。项目研制成果满足国家图书馆数字馆藏建设要求,符合数字对象资源的长期保存和使用目的。

本书即是以国家图书馆"文本数据加工标准与工作规范"项目的

研究成果为基础编写的，服务于国家图书馆数字资源建设，为国家图书馆文本数字对象加工提供原则性的指导。本书注重文本数字对象加工的全流程控制和管理，在满足数字资源长期保存和应用的基础上，根据文献载体特点和属性，推荐了文本数据的加工方法、格式标准、管理方式以及加工中对文献的保护方式，适用于国家图书馆文本数字资源的建设需要。图书馆等信息机构在信息化建设和数字资源加工实践中，也可以本书作为参考。

编者

2012 年 3 月 14 日

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第一部分　国家图书馆文本数据加工标准

1 引言

本标准规范是针对国家图书馆的实际情况，在保持与国家图书馆《数字资源（图像、音频、视频）加工标准与工作规范》《数字资源对象管理规范》及唯一标识符、元数据、长期保存等子项目组相关成果一致性的基础上，根据国家图书馆对象数据竞争性谈判项目第二包《文本数据加工标准与工作规范》研制业务需求书和成交合同而研制。

本标准规范服务于国家图书馆数字资源建设，针对国家图书馆的实际情况，兼顾目前国内外文本数据加工的发展趋势和需求，重点参考了美国国会图书馆国家数字图书馆计划的“关于文本和图像数据数字化转换的技术规范”和台湾地区“数位典藏国家型科技计划”的技术资料汇编以及科技部“中国数字图书馆标准与规范建设”项目“数字资源加工标准规范与操作指南”子项目的相关成果，本标准规范也将随着文本数据加工技术的发展而更新。若要以此报告为基础发展成文本数据加工的行业规范，需要在国家图书馆应用的基础上进行修改、补充和完善，并按照行业标准研制的正式程序组织有关单位共同研制。本标准规范仅对文本数据加工制定基本原则，具体细致的操作指导在本书“第二部分国家图书馆文本数据加工标准操作指南”中。

2 范围

本标准规范是针对一般文本文献的数据加工，规定的加工标准和工作规范。本标准规范适用于排版简单，以文字为主要表达形式，可

以存在少量图表的普通文献(不包括古籍善本、手稿等特殊文献)的文本数据加工制定。加工对象可以是一般纸质文本文献,也可以是印刷型文献经过数字转换后的图像文件。

本标准规范规定了国家图书馆文本数据加工的工作流程、文本数据制作、元数据加工、文件管理及质量管理方面的内容。

3 规范性引用文件

下列文件对于本标准的应用是必不可少的。凡是注日期的引用文件,仅所注日期的版本适用于本文件。凡是不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB 2312—1980 信息交换用汉字编码字符集—基本集(Code of Chinese Graphic Character Set for Information Interchange—Primary Set)

GB/T 2828.1—2003 ·计数抽样检验程序 第1部分:按接收质量限(AQL)检索的逐批检验抽样计划[Sampling Procedures for Inspection by Attributes—Part 1: Sampling Schemes Indexed by Acceptance Quality Limit(AQL)for Lot-by-lot Inspection]

GB 13000—2010 信息技术 通用多八位编码字符集(UCS)[ISO/IEC 10646:2003:Information Technology—Universal Multiple-Octet Coded Character Set(UCS)]

GB 18030—2005 信息技术 中文编码字符集(Information Technology—Chinese Coded Character Set)

GB/T 18793—2002 信息技术 可扩展置标语言(XML)1.0[Information Technology—Extensible Markup Language(XML)1.0]

GB/T 23286.1—2009 文献管理 长期保存的电子文档文件格式 第1部分:PDF1.4(PDF/A-1)的使用[ISO 19005—1:2005: Document Management—Electronic Document File Format for Long-term Preservation - Part 1: Use of PDF 1.4(PDF/A-1)]

GB/T 25100—2010 信息与文献 都柏林核心元数据元素集(Information and Documentation—The Dublin Core metadata element set)

ISO/IEC 10646:2003 Information Technology—Universal Multiple-Octet Coded Character Set(UCS)信息技术 通用多八位编码字符集(UCS)

ISO 32000—1:2008 Document Management—Portable Document Format—Part 1:PDF 1.7 文献管理 便携式文档格式 第一部分:PDF 1.7

4 术语和定义

下列术语和定义适用于本标准。

4.1 文献(Document)

记录有知识的一切载体。具体地讲,就是指用文字、符号、图像、声频、视频等手段,记录下来的人类知识的各种载体。文献具备三个要素:一是知识,即被固化于物质载体上的知识信息;二是记录,即文献的记录手段或方式;三是载体,即记录知识信息的物质载体。

注:来源于《文献著录总则》(GB 3792.1—83)和《情报与文献工作词汇基本述语》(GB 4894—85)。

此为试读,需要完整PDF请访问: www.ertongbook.com