甘克勤◎著

标准大数据实践

STANDARD BIG DATA IN ACTION

中国质检出版社
中国标准出版社

标准大数据实践

甘克勤　著

中国质检出版社
中国标准出版社
北　京

图书在版编目(CIP)数据

标准大数据实践/甘克勤著．—北京：中国标准出版社，2016.12
ISBN 978-7-5066-8488-0

Ⅰ.①标… Ⅱ.①甘… Ⅲ.①数据处理—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2016）第 285330 号

内 容 提 要

本书解读了当前大数据发展的基本背景，重点阐述国家标准馆信息化建设存在的问题，以及当前信息化建设要解决的 5 大问题：长期保存能力不足、数据找不到、数据找不全、数据找不准、标准情况说不清。本书围绕这 5 个问题，从业务和技术角度寻找解决办法，给出解决思路以及当前阶段性成果。

本书可供标准行业及文献行业的从业者参考。

中国质检出版社
中国标准出版社 出版发行
北京市朝阳区和平里西街甲 2 号（100029）
北京市西城区三里河北街 16 号（100045）
网址：www.spc.net.cn
总编室：（010）68533533 发行中心：（010）51780238
读者服务部：（010）68523946
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销

*

开本 710×1000 1/16 印张 9 字数 141 千字
2016 年 12 月第一版 2016 年 12 月第一次印刷

*

定价 35.00 元

前　言

2010年以来，随着国家标准馆采编流业务系统以及对内对外服务系统的搭建完成，存量文本资源图像化工作的结束，国家标准馆的信息化建设从基础设施建设期进入了深化发展的时期。与此同时，随着全社会信息化水平的飞速发展，“云计算”“互联网＋”“大数据”“物联网”等概念的不断更新，人们对用户体验的追求也在不断提升，对标准使用的诉求也越来越丰富，如在查找标准过程中，找不到、找不全、找不准，在分析标准过程中说不清等；另一方面，在日常具体业务开展和系统运行与扩展过程中，存在系统运行效率低下、数据共享手段不丰富、系统维护修改困难等现实问题。

本书总结了近年来国家标准馆信息化建设的实践，从标准信息服务社会需求和存在的问题表象出发深刻分析问题存在的原因，研究最新的大数据理论、方法和技术，将其与国家标准馆信息化建设的业务需求相结合，逐步形成了适用于标准信息服务的理论框架、方法体系和技术手段，最终完成了内部系统的软硬件架构优化，建成了多粒度（微观的指标与宏观的统计）、多终端（PC端与移动端）的标准大数据服务系统。

在作者看来，大的互联网概念离不开脚踏实地的工作落实，正如“互联网＋”落到实处就是“程序员＋”，通过程序员的工作，改变传统行业的工作状态。与此相同，“大数据”的内涵远远不是字面上的数据量大，更多的内涵则是数据的多样性、关联性和多维度统计，从而衍生出奇妙的组合，得到意想不到的结果。标准文献与

其他文献一样，其数据之大不是体现在文本数量上，而是体现在文本内容所包含的知识元。在标准中，更多的知识体现在标准化对象的指标上，针对标准文献内容的挖掘是本书的一个重点。另一方面，尝试针对标准进行多维度的统计分析，也是本书的一个方向。从“大数据”的外延来看，所有的基础建设都属于大数据实践的范畴，大到各种云平台（如“阿里云”“百度云”等），小到国家标准馆私有虚拟机的搭建和基础软件的改造，都在实践着“大数据”的理念，本书对此也进行了阐述。

本书从基层实践出发，注重解决国家标准馆信息化建设过程中遇到的实际问题，适用于标准和其他相似文献与信息服务的从业者作参考。

本人的专业是信息化技术，从事标准文献信息化工作六年时间。本人认为，标准文献信息化建设的最终目的是将标准转换为知识，标准检索系统变成具有人工智能内涵的专家系统，进而让标准化知识融入人们的日常生活，成为社会化大生产及人民日常生活不可或缺的一部分。本书只是展现了一个阶段性成果，标准文献信息化工作还任重道远，路漫漫其修远兮，吾将上下而求索，生命不息，奋斗不止。

著　者

2016 年 10 月

目　录

1 绪 论

本章从大数据研究的背景出发，列出当前标准文献信息化建设遇到的问题以及解决问题的思路与方向，为全书的内容组织梳理出思路。

1.1 背 景

数据资源已经成为全球经济发展不可或缺的组成部分，并逐渐成为一个和传统资本、实体资产同样重要的生产要素。数据的价值蕴含于对数据的分析和有效处理，形成从模型驱动的研究方法到数据驱动的研究方法的演进，这一变化正推动着传统计算科学向数据科学的转变，并对各行各业带来由基于经验的决策向基于数据的决策的升华。第六次科技革命将是改变人类对自己的认知，极大提高人类对自身掌控能力的交叉学科革命。人们将融合信息科学、社会科学、生命科学等，打破传统的学科界限和思维模式，重新打造各行各业，挖掘其价值。

1.1.1 大数据应用蓬勃发展

随着移动互联网、物联网、社交网络等技术与网络应用的兴起，全球范围内数据量迅猛增长，大数据时代已经来临。学术界和工业界都对大数据赋予大量的关注并开展了深刻的讨论。

2000 年，英国 e-Science 计划的启动；2003 年，英国超大级国家研究数据中心、国家级学科领域研究中心，数据服务中心提到日程；2007 年，英国发布《发展英国科研与创新信息化基础设施》，研究大数据中心的基础设施、技术和应用服务（数据和信息的产生数据的保存和管理，数据的查询和导航，虚拟研究团体，网络、计算和数据存储设施，数字版权管理等）；2007 年，

美国NSF发布了《21世纪的科研信息化基础设施》，提出了《国家数字化数据框架》；2012年3月29日，美国奥巴马政府发布了“Big Data Big Deal”，它是美国政府继“国家信息高速公路”（1992年）之后的又一重大战略计划，将大数据上升到国家战略。

大数据在全球范围内备受关注，对其定义也有多种提法。IBM提出3V，即认为大数据具备规模性（volume）、多样性（variety）和高速性（velocity）3个特征，规模性是指数据量巨大，达到TB级甚至PB级，多样性是指数据类型繁多，包括结构化数据和非结构化数据；高速性是指数据创建、处理和分析的速度持续在加快。在此基础上，有人提出了4V定义：大数据还应具有价值性（value），以及5V定义：大数据还应具有精确性（veracity）。除此之外，维基百科的定义认为大数据是指难以用常用软件工具在可容忍时间内抓取、管理以及处理的数据集。

IDC的研究报告称，到2020年全球数据使用量预计暴增44倍，达到35.2 ZB，即全球大概需要376亿个1 TB的硬盘来存储数据。企业中20%的数据是结构化的，同时80%是非结构化或半结构化的。此外，结构化数据增长率为32%，而非结构化数据增长约为63%，至2015年，非结构化数据占有比例达到互联网整个数据量的75%以上。在中国，“大数据”概念正在引领中国互联网行业新一轮的技术浪潮，2014年年底中国互联网行业持有的数量总量约为1.9 EB，而到2015年，这一数据已经增长到8.2 EB。

1.1.2 大数据技术不断更新

数据资源已经成为全球经济发展不可或缺的组成部分，并逐渐成为一个和传统资本、实体资产同样重要的生产要素。数据的价值蕴含于对数据的分析和有效处理，形成从模型驱动的研究方法到数据驱动的研究方法的演进，这一变化正推动着传统计算科学向数据科学的转变，并对各行各业带来由基于经验的决策向基于数据的决策的升华。大数据技术具体包括：

（1）云计算技术：它为大数据应用提供运行的基础环境，是最为重要的支撑技术，会涉及虚拟化技术、分布式数据存储、分布式计算、多租户管理、资源调度和优化、能耗管理等众多方面的技术。

（2）高性能计算技术：从体系结构上为一些大数据应用提供高性能计算的保障。例如，利用高性能计算或者一些先进的硬件技术支撑诸如高通量计算、大规模科学计算等特殊的大数据应用需求。在依赖大数据的深度学习应用背景下，高性能计算的支撑显得越发重要。

（3）高速、可信的网络技术：InfiniBand 和诸如 RDMA 的一些正在兴起的高速网络技术，也是大数据应用的重要支撑技术。另一方面，从信息安全的角度考虑，可信网络和可信计算在大数据背景下显得尤为重要。

（4）物联网技术：近年来电子商务和相应的物流行业在我国发展迅猛。物联网、RFID、机器产生的数据将成为大数据的重要来源，也是大数据的重要应用。与之相关的车联网、移动互联网等支撑技术都会成为这方面大数据技术的重要内容。尤其考虑到交通、环境、食品安全等方面的难题，我们迫切需要利用与大数据相关的物联网技术优化线下的物流环节等。

（5）自主可控的操作系统和信息安全技术：自主可控的操作系统是信息安全的基础。这不仅涉及大数据集群平台使用的操作系统安全，还包括海量的终端设备（如手机）所使用的操作系统的安全，以及众多互联网、移动互联网应用的信息安全等。在很多大数据的重大应用领域里，信息安全是战略层面的，可能比应用本身更为重要。从信息安全的角度看，相关支撑技术的自主可控意义非同一般。

（6）可信数据库技术：作为一类特殊和重要的系统软件，数据库在大数据背景下的挑战非常巨大。在很多诸如银行的重要应用领域，由于历史原因，数据库软件并没有很好地做到自主可控，这不仅仅让我们牺牲了很多经济利益，在信息安全方面也存在很大的风险。大数据背景下，给我国数据库技术的发展提供了新的机遇，但同时研发支撑大数据应用的、可信的、高性能的数据库技术也是非常棘手的。

1.1.3 标准信息化建设重要意义

标准化资源服务政府和社会的能力和水平急需创新发展。标准作为经济社会发展的重要技术支撑，已经成为国家治理体系和治理能力现代化的基础性制度。国际社会高度重视标准化工作，ISO 和 IEC 发布国际标准接近 3 万项，

美、德、英、法、日等国家纷纷将标准化工作提升到战略高度。经过多年发展，我国标准化工作取得大量成果，标准体系初步形成，国家、行业、地方标准接近 10 万项。大量的标准文献，是我国经济社会发展的重要技术资源。以此为基础，在国家的支持下，国家标准馆已经建设成为国内最大的标准文献馆藏中心，藏有 60 多个国家、70 多个国际和区域性标准化组织、450 多个专业协（学）会的成套标准以及全部的中国国家标准和行业标准，收集了160 多种国内外标准化期刊和 7000 多册标准化专著。随着“京津冀协同发展”“长江经济带建设”以及“城乡一体化”发展等战略和规划的落实推进，标准信息服务的需求日趋强烈。为此，《国家标准体系建设发展规划（2016—2020 年）》专门提出开展加强各级标准馆建设的任务要求，形成以国家标准馆为中心，辐射各地区、专业领域的各级标准馆布局，形成覆盖范围广、资源全面、设施齐全的标准信息储备基地，有助于实现社会各个层面及时获取技术标准资源，提升标准的有效性、更好地发挥标准的作用。

从战略意义上讲，《中国制造 2025》和《国家标准体系建设发展规划（2016—2020 年）》等重大战略和规划，都对开展标准信息服务提出了明确要求，将提升标准信息服务作为“十三五”时期加快完善国家标准体系，构建政府和市场共治的二元化新型标准体系的重要举措。不仅如此，随着我国标准化工作改革工作的不断深化，标准化面临着创新发展的新形势与新需求。通过开展标准大数据研究，将有助于加快落实《创新驱动发展战略》，推动新技术、新产品、新服务转化为技术标准，从源头上帮助提升技术标准的研制能力、在工作过程中帮助提升技术标准的技术水平和科学性、从末端强化标准化资源的有效分配与共享，增加技术标准的有效供给，提升标准化工作的质量和效益。

1.2 存在问题

随着大数据在社会各行各业的不断深化应用，特别是在搜索领域用户体验的不断优化提升，公众对于标准文献信息获取的需求也不断提升。

1.2.1 数据问题

1.2.1.1 标准长期保存瓶颈

标准文献既是国家重要的基础信息资源，也是国家战略性公共科技资源，已经与专利文献、科技文献一同列入《国家中长期科学和技术发展规划纲要(2006—2020年)》。标准战略成为国家实施科技强国的三大战略之一。实施技术标准是创新成果产业化的关键环节，是实现跨越式发展必不可少的要素，标准文献的持续获取和有效利用是实施国家标准战略的首要环节。

为支撑我国不断深化推进的“标准联通一带一路”战略，国家标准化管理委员会将“开展大宗进出口商品标准比对分析”列为主要工作之一，“开展沿线重点国家技术法规和标准信息收集、标准翻译，开展优先领域大宗商品标准比对分析，完成优先领域大宗进出口商品标准比对分析研究报告”等工作则是该项工作落地的举措。为落实国家标准化管理委员会的重点工作，国家标准馆着眼于技术法规和标准信息的收集工作，已将“一带一路”沿线国家的标准列入标准采购计划并已经完成越南、新加坡等东盟国家，沙特、伊拉克等西亚国家，印度等南亚国家，以及南非、巴西等金砖国家，埃及、苏丹等非洲各国标准的采购，新增的采购资源以电子资源为主，纸本资源为辅。另一方面，传统品种持续的续订，使得国家标准馆馆藏资源逐年快速增长的趋势日益显著，对数字资源长期保存能力的要求不断提高。

按照国家标准馆的业务划分，标准文献资源的长期保存分为以下几个方面：纸本资源电子化与长期保存、外购原始光盘资源的长期保存、清洗整合后数据资源长期保存、深入挖掘的内容信息长期保存。目前，由于国家标准馆基础设施所限，只针对纸本资源电子化以及整合后数据资源进行了长期保存，缺乏针对外购的原始光盘数据和深入挖掘的内容信息长期保存的能力和机制。

国家标准馆数字资源长期保存的需求持续增长，对国家标准馆长期保存能力提出了新的挑战，主要体现在以下几个方面。

(1) 清洗整合后的数据持续增长

清洗整合后的标准数据是国家标准馆对外提供服务的主要来源，随着每年采集的标准不断清洗、整合、入库，标准数据资源逐年递增；同时，随着

采集范围的扩大，标准数据资源的增长量也呈现逐年递增的趋势。近3年，新增的标准容量分别是22.9G、38.3G和56.1G。加速增长的标准数据资源，使得我们必须重新审视标准文献长期保存的存储规模，以满足至少未来5年的标准数据资源爆发性增长。

（2）外购光盘资源长期保存需求凸显

光学介质的存储媒体具有易磨损性与不稳定性，使其有效保存期远低于纸介质和微缩胶片，如光盘表面在经过一段时间的读取后会出现划痕和磨损，很可能造成无法打开或部分内容无法读取。国家标准馆早期采购的部分标准光盘，目前就已经无法打开使用。目前，国家标准馆外购的标准文献电子资源的介质以光盘为主，由于光盘介质寿命有限和损坏无法恢复的特点，原始数据存储于光盘介质中，极不符合资源长期保存的基本要求，宝贵的标准文献资源存在“一经损坏，永久丢失”的风险。

与光盘介质不同，磁盘阵列是基于硬磁盘的存储方式。这些存储设备独立，通过光纤交换机与前端应用系统连接起来，形成一个光纤通信网络，在这种数据存储方案中，数据以集中的方式进行存储，加强了数据的可管理性；同时，具备可扩展性的通用数据共享同一存储池，降低了存储系统的总拥有成本。另外，磁盘阵列通过内置的硬件冗余和RAID保护可确保数据能够快速而准确地恢复，极大地增加了数据安全性，降低了数据丢失的风险。因此，磁盘阵列是数字资源长期保存的最优选择，目前已经成为主流的解决方案。

另一方面，外购标准原始数据的分散存储不利于标准信息资源的统一管理和溯源，同时数据清洗整合工作需要人工参与，周期较长。因此，国家标准馆对外购标准原始数据的统一存储，不仅能对标准电子资源进行有效的统一管理和溯源，更能极大地提升标准信息资源获取的效率，缩短国家标准馆从信息采集到信息服务的周期。

因此，将外购的标准文献资源的原始数据进行统一存储是国家标准馆数字资源长期保存的必要前提和重要手段，而原始数据的统一存储也对存储规模提出了新的需求。

（3）知识挖掘内容保存需求凸显

标准作为“为了在一定的范围内获得最佳秩序，经协商一致制定并由公

认机构批准，共同使用的和重复使用的一种规范性文件”，以科学、技术和经验的综合成果为基础，以促进最佳的共同效益为目的，是人类科学发展成果的重要展示方式之一。国家标准馆长期保存的不仅是标准文献本身，更是标准文献中包含的知识，对于科技创新起着重要的作用。

针对标准文献的长期保存，不仅需要进一步增加存储的规模，更需要增加数据挖掘的存储能力，存储标准文献及其包含的重要指标的关系网格，更好地为科研和工业生产提供有效的基础数据支撑。

海量电子形式的标准资源对国家标准馆信息长期保存的基础设施提出了新的迫切的需求，近两年国家标准馆经多方筹措资金，升级了自身的基础硬件环境，包括存储扩容、应用服务器升级、虚拟化改造。然而，针对国家标准馆数据资源的长期保存需求，改造后的基础能力仍显得不足，主要体现在以下几个方面。

①数据存储能力已趋于饱和

目前，文献资源统一存储规模为 40T，已经占用空间 35T，主要用于标准电子资源的存储、数据库存储、各类应用服务器空间等。随着采购标准资源每年的增长，现有存储空间只能满足未来 1～2 年的需要。急需购置新的存储资源，以确保我国宝贵的标准文献战略资源得到妥善的长期保存。

②原始外购光盘统一存储能力缺失

目前，原始外购的标准光盘以分散保存为主，国家标准馆基础设施缺乏统一存储的能力。随着时间的推移，原始外购光盘损坏的风险越来越高，急需针对原始外购光盘建立起有效、统一的长期保存机制。

③标准内容长期保存能力不足

标准文献知识挖掘主要是指针对非结构化的标准文本，基于结构化和指标化的技术工具挖掘标准文本中的内容指标，基于分类和聚类的技术工具对标准中的内容指标进行重新组织，进而为政府、企业提供基于标准文献知识挖掘的创新型知识服务，如中外标准比对、标准段落检索、标准指标检索、标准情报预警、标准大数据分析等。

目前，国家标准馆针对标准内容的 OCR 识别、结构化拆分、指标挖掘工作逐步开展，其中的过程数据与成果文件同样需要大量的存储空间。

1.2.1.2 标准找不到

2000～2010年，国家标准馆完成了首轮信息化建设，实现了纸质图书馆向数字化和信息化的跨越，基本告别了卡片检索和排架提书的时代，进入了数字图书馆时代。然而，随着国家标准馆不断扩大采集资源的范围和数量，包括一些以前很少采集的小国家的标准资源，如东南亚、沙特、印度等国的标准，导致了馆藏标准快速增长。传统的字段映射、拼接、替换、修改的数据整合方式，难以应对多品种、大批量的馆藏量增长，多个不同来源的标准文献数据分散存放在各自系统中，甚至分散存放在各自光盘里，“标准找不到”成为读者和对外服务的咨询馆员最常见的抱怨话题。

从技术上角度看，大数据时代异构数据的集成是连接数据孤岛的重要内容，也是近年来持续热门的研究方向。作为馆藏机构，国家标准馆内部也存在多源异构数据的问题，主要体现在来自不同国家、不同数据结构文献的集成。传统人工编制映射规则将异构变成同构的做法，难以满足大数据的集成需求。而这个课题已经成为传统的标准图书馆必须解决的当务之急，即“有馆藏找不到”必须解决。解决这一问题的前提正是前文所述的“充足的存储空间和统一的物理存储”，进而应用大数据相关索引技术，实现“多源异构数据”的集成检索。解决“标准找不到”这一看似容易，实则工作量较大的顽疾。

1.2.1.3 标准找不全

传统的标准题录检索（主要针对标准号、标准题名、适用范围、主题词）相较传统的卡片管理实现了历史性的跨越，然而，仅靠题录字段难以完整的概括标准全部信息，进而造成检索的不够完全，检索的查全与关联一直是图书馆学、互联网技术关注的重点，标准文献的查全主要依靠内容的有效检索命中和知识关联，本书将阐述国家标准馆在标准全文索引和知识关联（引文关联、文本相似度）的研究与应用进展，力图为用户提供更好的查全保证。

1.2.1.4 标准找不准

当今世界处在一个“数据爆炸”的时代，由于数据的长期积累和存储技术的不断发展，电子数据量变得非常庞大，目前数百万乃至上千万条记录的数据库不罕见，国家标准馆的馆藏标准记录正是这样的上百万条记录。随着Internet的迅速普及，人们可以轻易获取大量的数据，但是要从数据中获取真