人类遗传资源样本信息元数据质量评价模型探讨

钟 凯[1] 赵 伟[1] 宋立荣[2]
（1.中国科学技术信息研究所，北京 100038；2.北京联合大学，北京 100191）

摘要：人类遗传资源是国家重要的战略资源。当前我国对于该资源的建设重点集中在样本库建设上，而样本信息库的建设及其质量控制的实施较为薄弱。文章通过研究已有人类遗传资源样本信息管理相关成果及元数据质量评价在其他领域的应用，提出适用于人类遗传资源样本信息的元数据质量评价模型。模型由评价维度的形式质量、内容质量和效用质量3个层面，功能维度的资源控制、资源描述和资源效用3个方面，以及重要性维度的核心层、重要层和扩展层3个层级共同构成，为将来开展人类遗传资源样本信息元数据质量控制工作奠定理论基础。

关键词：人类遗传资源；样本信息；元数据；质量评价；元数据质量
中图分类号：G203 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2018.04.010

Discussion on Metadata Quality Evaluation Model of Human Genetic Resources Sample Information

ZHONG Kai[1], ZHAO Wei[1], SONG Lirong[2]
(1.Institute of Scientific and Technical Information of China, Beijing 100038; 2.Beijing Union University, Beijing 100191)

Abstract: human Genetic Resources are important national strategic resources. The focus of the construction is the biobank construction, while the construction of database of sample information and the implementation of quality control are relatively weak. Based on the research of the achievements of the existing sample information management of human genetic resources and the application of metadata quality evaluation in other fields, we propose a metadata quality evaluation model that is applicable to sample information of human genetic resources. The model consists of three dimension aspects: form quality, content quality and utility quality; three function aspects: resource control, resource description and resource utility; and three importance aspects: core, key and extension. This could lay the theoretical foundation for metadata quality control of sample information in the future.

Keywords: human genetic resources, sample information, metadata, quality evaluation, metadata quality

0 引言

人类遗传资源样本信息是具有重要价值，并能为生命医药领域基础理论探索和实践应用两方面奠定研究基础，且涉及国家安全的重要战略数据。1998 年国务院办公厅转发施行的《人类遗传资源管理暂行办法》[1]中规定了人类遗传资源是指含有人体基因组、基因及其产物的器官、组

作者简介：钟凯（1992—），男，中国科学技术信息研究所硕士研究生，研究方向：信息资源开放与共享；**赵伟**（1975—），女，中国科学技术信息研究所研究员，博士，研究方向：科技资源管理（通讯作者）；**宋立荣**（1971—），男，北京联合大学研究员，博士，研究方向：科技信息共享、信息质量。

基金项目：国家重点研发计划“中国人类遗传资源样本库建设”（2016YFC1201700）；国家社会科学基金项目“网络环境下科技信息资源建设中的质量元数据及评估应用研究”（12BTQ016）。
收稿时间：2018 年 2 月 12 日。

织、细胞、血液、制备物、重组脱氧核糖核酸（DNA）构建体等遗传材料及相关的信息资料。

当前，我国已经建设了众多规模不一的、涉及人类遗传资源的样本库，但遗传资源样本信息开放共享的程度和利用率却不高。遗传资源样本信息能揭示实体资源的基本信息，为用户寻找样本资源提供依据，促进海量人类遗传资源样本信息的开放共享，极大地使样本资源得到充分利用，满足社会对该种具有不可再生性的资源的迫切需求，尽可能地发挥其作用与价值。

国家科技基础条件平台将科学标本资源分为八大类，其中一类就是人类遗传资源，这表明人类遗传资源作为科学标本资源，既有所有科学标本所具有的共性描述信息，也有其特性描述信息。本研究侧重于人类遗传资源样本信息库元数据质量的评价。而人类遗传资源样本信息库来源于人类遗传资源样本库，这使得人类遗传资源样本信息元数据也有区别于其他元数据的特殊之处。人类遗传资源样本库由样本实体库和样本信息库共同构成，信息库侧重于信息的存储，即主要存储与实体资源样本相关联的数据、信息。孙晓东等[2]认为信息库数据不仅应包括样本实体信息，还应包括样本关联病史资料、所有医疗记录等。涉及病史资料、医疗记录，是多数人类遗传资源与其他资源所不同的特点。长文本的内容是全文存储还是进行关键词存储、关键词是人工选取还是用基于机器学习的算法选取又影响着科研人员后期对资源情况的判断。此外，人类遗传资源样本信息还具有涉及伦理的信息，这些特殊元数据的评价是目前已有元数据质量评价模型尚未涉及的。

人类遗传资源样本信息元数据的特殊性还在于其描述的实体是实物，不同于信息资源元数据，因为信息资源元数据与资源实体是一体的。传统的元数据具有“所见即所得”的特性，而样本实物的元数据与样本实体是分离的，不具有内容方面的客观描述。

基于对众多文献归类分析，本文将“样本信息资源”定义为样本信息及其衍生、相关信息的集合，包括了两大类数据：一类是样本数据库及其内容数据，样本源相关的病例信息、随访信息等临床和实验相关信息；另一类是样本信息平台的信息，包括样本相关的原始数据和汇交数据时填写的样本信息元数据。

1　相关研究基础

1.1　人类遗传资源样本库

根据对文献的归纳整理，从学者的研究来看，人类遗传资源样本库包括信息标准建设和信息共享平台建设两个方面。

（1）信息标准建设

在信息标准建设方面，基于促进生物样本库共享的目的，Norlin等[3]为了促进泛欧洲生物样本库与生物分子资源研究中心（Biobanking and BioMolecular resources Research Infrastructure，BBMRI）的数据共享，使用户通过标准化数据格式访问BBMRI的样本信息资源，对BBMRI的最小元数据标准MIABIS（Minimum Information About BIobank data Sharing）进行了分析，并结合国外多家样本库元数据标准如MedicineNet.com、EMBL（EFO）、P3G等提出了2.0版本的数据标准，并列举了样本库和信息库之间的属性关系。Quinlan等[4]研究了基于MIABIS的面向使用的集成生物样本库网络数据标准，提出了多种信息共享服务的元数据标准如最小数据集、最佳实践标准等。

李怡等[5]认为生物样本信息资源库由生物样本、临床/病理等表型信息和资源管理数据库三部分构成，研究了国内外生物样本信息资源库的发展和现状，并指出了当前阶段我国在生物样本信息资源库建设和管理中的主要问题有：相关法律法规不健全和没有统一的标准操作流程和质量控制体系。

（2）信息共享平台建设

国际生物和环境样本库协会（ISBER）[6]作为解决与生物和环境样本库相关的科学、技术、法律和伦理问题唯一的全球论坛，提供了实践建议。同时，每年也会召开大型学术会议，召集全

球样本库领域的科研人员共同商讨样本库发展之路。BBMRI作为欧洲地区最大的生物样本库资源网络，其结构为分布式枢纽，协调各项活动，在其首页上提供给不同类型人员多种查询信息的渠道。

2003年7月，根据《国家中长期科学和技术发展规划纲要》，中国人类遗传资源平台（National Infrastructure of Chinese Genetic Resources，NICGR）作为国家科技基础条件平台的重要组成部分，同步启动建设，并于2007年投入使用[7]。表1列举了国内外部分较知名且具有代表性的人类遗传资源样本信息库。

由此可见，人类遗传资源样本信息库的研究目前还处在初级阶段，目前主要以搭建样本信息共享平台为短期目标，而对于平台数据内容质量的管理尚未开展。

1.2　元数据质量评价

对于元数据质量，国内外目前仍尚无统一的定义。美国学者Bruce、Hillman[8]提出元数据质量的概念并将其定义为“元数据满足需求和目标的程度”。其他学者结合各自研究的对象则有不同的阐述。不论哪种元数据质量的定义，都是评价元数据满足某种需求的程度[9]。

元数据质量的评价方法，按评价结果的呈现方式划分有：定性评价、定量评价和定性定量相结合等方法；按评价过程划分有：人工评价和统计评价的方式（表2）。在表2中直观地表明，人工对元数据质量评价和使用统计方法对元数据质量进行评价的评价能力相去甚远，完全不在一个数量级上。随着信息技术的发展，运用统计学习的方法对信息资源进行质量评价能充分体现计算机辅助评价的优势。林爱群[10]在对机构知识库中自动生成的元数据进行质量控制的研究时主要参考了国外Bruce和Hillman的元数据质量模型提出了评价的两个指标：完整性和精确性，但其研究没有涉及评估的目的、方法，且仅有两个指标的评价体系也不够完整。

元数据质量评价体系尚在探索阶段，目前已有的元数据质量评价模型，如：Moen等提出的评估模型、Stvilia模型、Bruce & Hillman模型以及

表1　国内外部分人类遗传资源样本信息库

样本库名称	地点
国际生物和环境样本库协会（ISBER）	加拿大
泛欧洲生物样本库与生物分子资源研究中心（BBMRI）	欧盟
公共人口基因组计划（Public Population Project in Genomics，P3G)	加拿大
中国人类遗传资源平台	中国
国家基因库	
北京重大疾病临床数据和样本库	

表2　不同元数据质量评价方法

研究者	方法	评价数据/（条/记录数）	评价视角
Greenberg等	人工	11	非专家定义的元数据
Moen等	人工	80	所有实例质量
Wilson	人工	100	非专家定义的元数据
Shreeves等	人工	140	所有实例质量
Stvilia等	人工	150	标识质量问题
宋立荣等	人工	150	用户服务需求
喻乒乒 等	人工	–	书目数据质量
Najjar等	统计	3700	元数据标准的使用
Hughes	统计	27000	实例完整性
Bui和Park	统计	1040034	实例完整性

注：整理自多篇相关文献。

黄莺提出的双维四核心模型、蒋引娣提出的元数据质量评价模型。各模型的优缺点归纳总结如下。

Moen模型基于GILS元数据的评估结果而非仅仅针对元数据质量，从用户、政策、技术、内容和标准5个方面展开，提出了21个指标的评估模型[11]。Stvilia等提出的质量评价模型得到了非常广泛的研究应用，归因于其模型具有良好的理论基础和建立思路。但是，Stvilia模型指标间存在重复且其含义不同，利用此模型开展元数据质量评价过程较为复杂，需要根据不同环境明确重复指标的不同含义，对于实现自动化评估也会产生相关限制。

Bruce和Hillman在Stvilia的模型基础上，对其进行优化，使得元数据质量评价模型脱离了元数据的创建、应用环境，具有较广的适用范围。蒋引娣从经济学的角度提出了具有创新性的质量评价模型，但是在实际运用中缺乏一定的可操作性。

黄莺学者提出的双层四核心元数据评价模型是为数不多的针对元数据质量评估而建立的模型，通过分析早期的Moen模型、Stvilia模型、Bruce & Hillman模型，提出了具有可扩展性和实用性特征的由影响元数据质量的核心要素和可选要素共同构成的双层结构模型。总的来说，为了扩大元数据质量评价模型的使用范围，摒弃了与元数据开发环境、应用环境相关的评价维度，仅保留了评价元数据自身质量的评价维度，因此具有较高的适用性。但是当运用到具体的领域中时，只有进行比较多的修改才能提高其模型的评价完整度。

总体上说，国内外对元数据质量评价的研究注重于结合具体领域进行实证分析，有关人类遗传资源样本信息元数据质量评价的研究较少，已有的成果也没有凸显人类遗传资源样本信息的特殊性。

1.3　人类遗传资源样本信息元数据评价要素

对于样本信息库而言，张育军对院级样本库的信息化建设进行了分析，实现全流程可视化管理。俞红、李海欣等从生物样本库的信息管理系统建设和管理进行了理论和实践研究，设计了单一病种样本库的信息管理系统，提高实体样本库使用效率。由此可见，各机构都已开展了实体样本库的信息管理工作，但是对信息管理工作所涉及的元数据尚未有考核评价模型。

目前，人类遗传资源样本信息库主要存在元数据标准不统一、元数据填写率较低、缺少评价标准等问题。因此，基于人类遗传资源样本信息平台的元数据标准，应研究如何提升元数据质量。样本信息元数据首先需要保证样本信息的管理需求，如对样本信息元数据进行宏观把控、对样本信息质量开展统计分析、对某一机构所上传的数据进行整体评价等，从中发现改进方向，不断提高元数据填写质量。在提倡资源开放共享的今天，人类遗传资源样本信息平台的功能之一就是给科研人员提供样本信息及获取渠道，因此也需要考虑到用户对于样本信息检索时的需求。最后，利用海量的样本信息元数据进行语义关联、深度挖掘研究，挖掘数据背后更深层次的科研价值，充分利用样本信息元数据。

元数据质量评价的关键是确定质量评价的维度，从管理者、用户的角度进行评价，从样本信息元数据的质量特性进行评价等。通过对各领域元数据质量评价文献提出的元数据质量评价维度进行梳理和归类，发现目前对于元数据的质量评价主要围绕完整性、准确性、一致性3个维度展开。人类遗传资源样本信息库元数据包含多个方面，综合元数据标准、样本信息的特性及信息库实际管理和业务需求，结合文献资源和实际样本信息平台实践做法，本文统计、分析、提炼了10个方面的样本信息库元数据质量维度。

通过专家访谈，进行了维度重要性的打分，经过维度调整，最终将质量评价维度演变成可测评的下列7个元数据质量维度：一致性、规范性、准确性、安全性、完整性、及时性、可获得性。

2　元数据质量评价模型的提出

2.1　评价基本思路

样本信息库元数据质量评价指标选择过程如

图1所示，这是开展评价维度筛选、确定工作的前提。整体评价思路由顶层质量目标为起点进行逐层分解，依次分析质量需求、质量准则、质量维度和指标层以及评估层的内容。

质量评价目标来源于人类遗传资源样本信息质量要求，根据文献调查和现场调研的结果，海量的人类遗传资源样本信息面临着质量提升的压力。因此，元数据质量评价目标一方面结合样本信息平台提出可管、可控、可溯源的质量管理目标，另一方面提高用户在使用人类遗传资源样本信息库时的满意程度。

对于质量需求的分析，依据元数据应用的环节不同，质量需求可分为资源描述质量需求、管理功能质量需求和资源使用质量需求。其中，资源描述质量需求对应资源描述质量准则，侧重于内容质量层面的评价；管理功能质量需求对应管理功能质量准则，侧重于形式质量层面的评价；资源使用质量需求对应于资源使用质量准侧，侧重于效用层面的评价。因此，对于不同准则的评价，通过多方面的整合，筛选出评价的质量维度，并最终形成完整的评价体系。

2.2　评价模型构建

人类遗传资源样本信息元数据质量评价模型由3个质量层面、3个需求方面以及3个层级构成（图2）。其中，3个层面是指从评价维度进行区分，分成形式质量、内容质量和效用质量3个

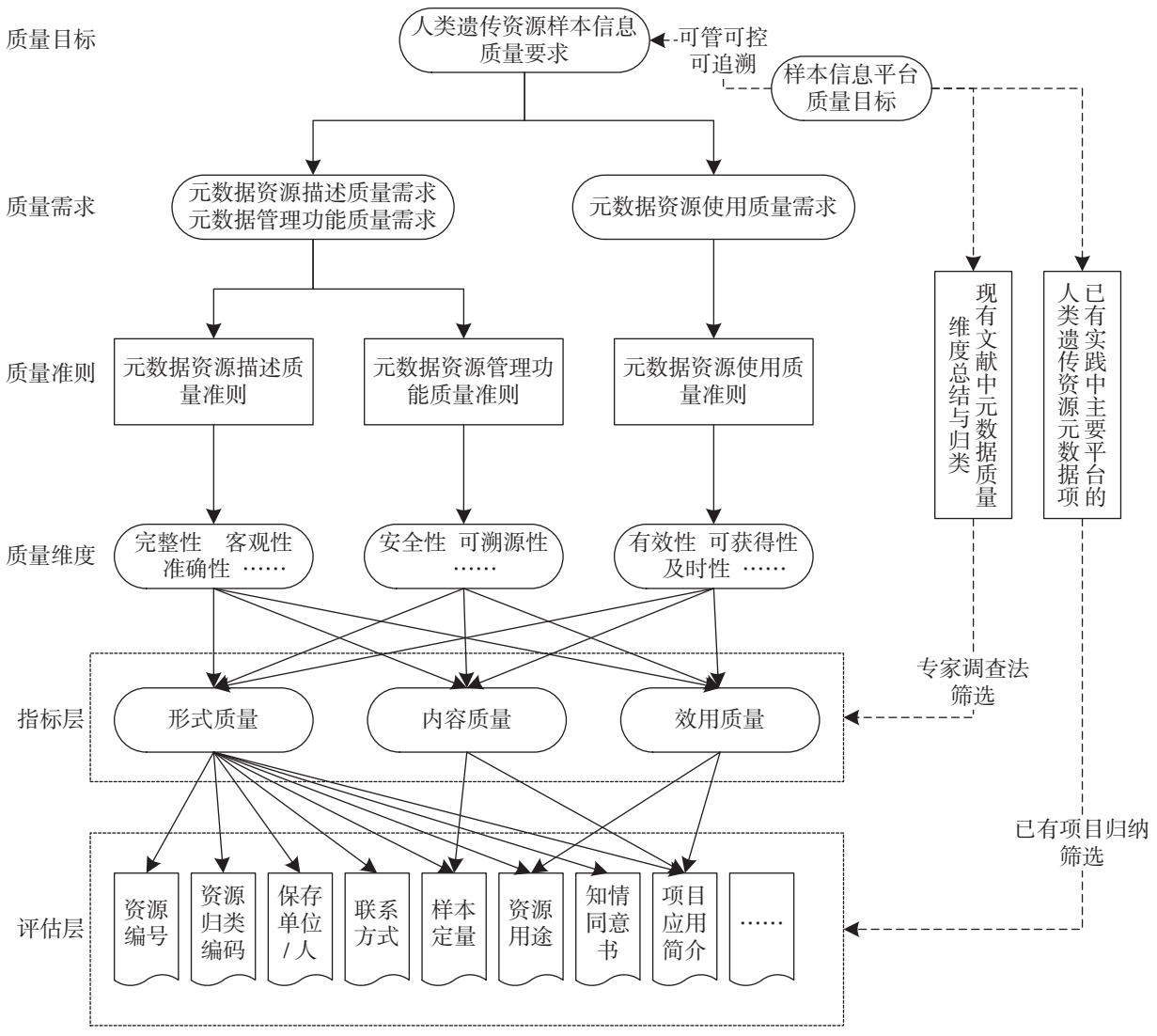

图1　元数据质量评价指标选择示意图

层面。各个质量层面又侧重于不同的质量准则方面：形式质量侧重于资源控制方面，内容质量侧重于资源描述方面，效用质量侧重于资源效用方面。最后，从质量维度的重要程度进行划分，分为核心层、重要层和扩展层 3 个层级。

就人类遗传资源样本信息库领域而言，本文界定了评价目标的 3 层需求方面：一是资源描述方面。搭建人类遗传资源样本信息数据的收录平台，研究海量人类遗传样本库数据资源管理技术，主要目的之一是为了对我国海量人类遗传资源及其信息进行统一管理、存储。元数据内容是否符合标准描述规范，在元数据内容形式上按照元数据标准进行填写，同时能完成对样本信息的基本描述作用即可。二是资源管理方面。基于文献和相关管理规范的基础上，对于样本信息库基本管理功能的理解为达到信息库的基本需求即对海量样本信息以元数据方式进行存储，保证数据存在、格式/逻辑符合正常要求即可，同时对样本资源的变化须有文字的记录。三是资源使用方面。搭建平台的另一主要目的是为了促进我国人类遗传资源的共享交流，避免科研人员因寻找合适的样本资源浪费大量的时间、人力、金钱等，同时也能避免重复地制作同一需求的样本或相似度极大的研究，通过信息平台搜索符合研究目的的样本资源，将极大地提高生物/生命科学领域的研究效率。另外，将海量的人类遗传资源样本信息进行关联，运用数据挖掘等方法发现一些潜在的关联信息，这是项目提出的一个建设要求。而开展数据挖掘工作需要大量的相关信息和有效的高质量的元数据。

资源描述、资源管理和资源使用 3 个方面的需求基本涵盖了人类遗传资源样本信息元数据在实际应用过程中可能涉及的需求，考虑了资源的提供方、资源的管理方以及资源使用者 3 种人群的需要，使得所构建的模型能够应用于实际操作中。同时，根据 3 方面的需求倒推了各自所需的服务，使得模型整体既能从管理者的角度出发开展人类遗传资源样本信息元数据的质量评价，又能兼顾数据提供者和数据使用者的切身需求。

3　结语

本文在分析已有元数据质量模型的基础上，

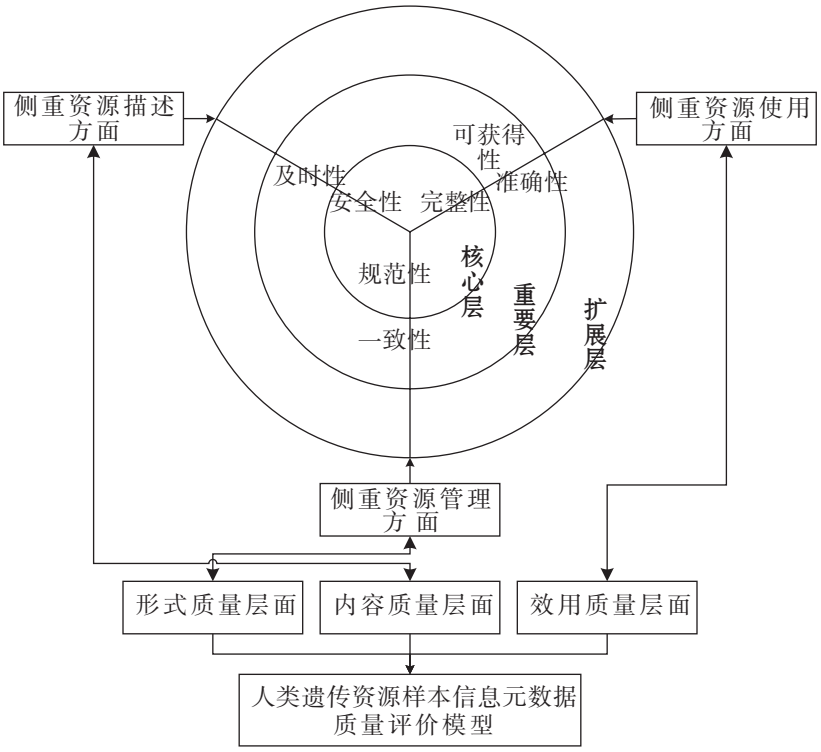

图 2　样本信息元数据质量评价模型

发现已有元数据质量评价模型并不适用于人类遗传资源样本信息的元数据质量评价，究其原因是已有模型以一些共性特征为基础所建立的，因此在结合人类遗传资源样本信息的相关特征包括医疗记录、病史资料等长文本内容以及伦理信息等特性数据后，提出了综合资源描述、资源管理和资源使用3个视角的人类遗传资源样本信息元数据质量评价模型。通过3个层面、3个方面和3个层级的一一对应，将涉及的所有评价要素进行关联，尽可能地覆盖样本信息元数据会面临的各种应用情景。

在今后的研究中，将基于本文所提出的模型实现具体的人类遗传资源样本信息元数据质量评价体系和评价方法，尤其是验证人类遗传资源样本信息特殊字段的元数据质量评价，并根据实证分析结果提出对人类遗传资源样本信息平台的元数据管理的对策建议。

参考文献

[1] 科技部. 人类遗传资源管理暂行办法[EB/OL]. (2008−11−16)[2017−07−16]. http: //most.gov.cn/fggw/xzfg/200811/t20081106_64877.htm.

[2] 孙晓东, 朱鸿. 重视精准医疗在眼科临床实践中的应用[J]. 中华眼科杂志, 2016, 52(2): 85−88.

[3] NORLIN L, FRANSSON M N, ERIKSSON M, et al. A minimum data set for sharing biobank samples, information, and aata: MIABIS[J]. Biopreservation & Biobanking, 2012, 10(4): 343−348.

[4] QUINLAN P R, MISTRY G, BULLBECK H, et al. A data standard for sourcing fit−for−purpose biological samples in an integrated virtual network of biobanks.[J]. Biopreservation & Biobanking, 2014, 12(3): 184−191.

[5] 李怡, 张换敬, 刘国彦, 等. 生物样本信息资源库的国内外研究进展[J]. 中国医药生物技术, 2012(5): 369−372.

[6] ISBER. About ISBER[EB/OL]. (2017−09−13)[2017−07−01]. http: //www.isber.org/?page=About.

[7] 中国人类遗传资源平台. 中国人类遗传资源平台[EB/OL]. (2015−12−30)[2017−09−02]. http://www.egene.org.cn/nipcgr/index.jsp.

[8] BRUCE T R, HILLMANN D I. The continuum of metadata quality: Defining, expressing, exploiting[J]. Metadata in Practice, 2004: 238−256.

[9] NISO. A framework of guidance for building good digital collections−3rd edition[EB/OL]. (2007−12−01)[2017−01−15]. http: //www.niso.org/publications/rp/framework3.pdf.

[10] 林爱群. 机构知识库元数据的自动生成与评估研究[J]. 图书馆学研究, 2009(7): 21−23.

[11] MOEN W E, STEWART E L, MCCLURE C R. Assessing metadata quality: Findings and methodological considerations from an evaluation of the U.S. government information locator service (GILS) [C]. IEEE Forum on Reasearch and Technology Advances in Digital Libraries, Santa Barbara, California, USA, 1998.

（上接第5页）

[6] 孙平. 科研诚信的挑战与应对策略:记第二届世界科研诚信大会[J]. 科技管理研究,2011,31(22):219−222.

[7] 马佰莲,谢婧. 近十年国内科研诚信研究述评[J]. 齐鲁师范学院学报,2012,27(6):49−54.

[8] 陈雨,李晨英,赵勇. 国内外科研诚信的内涵演进及其研究热点分析[J]. 中国科学基金,2017,31(4):396−404. DOI:10.16262/j.cnki.1000−8217.2017.04.017.

[9] 杨东占. 构建信用体系 加强科研诚信制度建设[J]. 中国高校科技,2014(9):11−15. DOI:10.16209/j.cnki.cust.2014.09.037.

[10] 陈德春. 丹麦科研诚信建设及经验分析[J]. 全球科技经济瞭望,2016,31(11):24−27.

[11] 程如烟,文玲艺. 主要国家加强科研诚信建设的做法及对我国的启示[J].世界科技研究与发展,2013,35 (1): 153−156.

[12] 淮孟姣,潘云涛,袁军鹏. 美国科研诚信管理体系建设研究:以美国科研诚信办公室为例[J]. 全球科技经济瞭望,2016,31(12):8−13.

[13] 王涛,夏秀芹,洪真裁. 澳大利亚科研管理和监督的体系、特点及启示[J]. 国家教育行政学院学报, 2014(11):85−90.

[14] 蒯强. 法国倡导科研诚信和反对学术不端行为的举措[J]. 复旦教育论坛,2007(5):81−84.

[15] 王飞. 德国科研不端治理体系建设的最新进展及启示[J]. 中国高校科技,2017(5):11−14.DOI:10.16209/j.cnki.cust.2017.05.003.