遗传资源样本信息标识符的可溯源研究

曹 祺[1] 赵 伟[1] 何思远[2]
（1.中国科学技术信息研究所，北京 100038；2.武汉大学国际软件学院，湖北武汉 430000）

摘要：中国人类遗传资源平台是我国人类遗传资源的整合与共享平台，其服务于民族种群结构研究、健康与疾病研究等人类学、民族学、医学、生理学的学科研究，提供相应的遗传信息资源。从情报学的角度开展中国人类遗传资源平台样本信息的统一资源标识符版本管理研究，旨在提供一种即便于快捷资源管理，又能够实现对资源信息标识可追溯性的解决方案。收集整理了近年来同领域的研究成果，分析现有问题和不足，创新性地提出基于DOI协议和SVN溯源管理思想的标识符版本管理策略，并对研究者资源查找方法进行简要描述，为情报管理研究领域资源唯一标识的可追溯性理论完善提供借鉴。

关键词：中国人类遗传资源；DOI；标识符；版本管理；可溯源
中图分类号：G354.2 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.04.004

Research on the Traceability of Genetic Resource Sample Information Identifiers

CAO Qi[1], ZHAO Wei[1], HE Siyuan[2]
(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. Wuhan University International School of Software, Wuhan 430000)

Abstract: National infrastructure of Chinese genetic resources (NICGR) is a platform that integrates and shares human genetic resources in China. It is also an important research and application project in the 13th Five-Year National Innovation Plan of China. It serves the researches in national population structure, health and diseases, anthropology, ethnology, medicine and physiology, while providing corresponding genetic information resources. This paper studies the uniform resource identifier (URI) version management of the NICGR sample information from the perspective of information science, hoping to provide a solution that not only facilitates resource management but also achieves the traceability of resource information identifiers. This paper collects the research results in the same field of study in recent years, expounds the existing problems and shortcomings, puts forward the identifier version management strategy based on DOI protocol and SVN traceability management philosophy. It also briefly introduces the resource search methods for researchers. This study can provide a basis for perfecting the theories in traceability of resource unique identifiers in the information management research field.

Keywords: Chinese human genetic resources, DOI, identifier, version management, traceability

作者简介：曹祺（1988—），男，中国科学技术信息研究所博士后，工程师，主要研究方向：科技资源共享、评价与管理、科技人才信息分析；**赵伟**（1975—），女，中国科学技术信息研究所研究员，主要研究方向：科技资源共享、评价与管理、科技人才信息分析、生态环境管理（通讯作者）；**何思远**（1989—），男，武汉大学国际软件学院硕士研究生，主要研究方向：软件工程，数据挖掘。

基金项目：国家重点研发计划生物安全关键技术研发重点专项“中国人类遗传资源样本库建设”（2016YFC1201700）。
收稿时间：2017 年 5 月 19 日。

1　引言

2016年7月，国务院发布的《“十三五”国家科技创新规划》[1]提出了一项重要要求，即“推进人类遗传资源的系统整合与深度利用研究，构建国家战略生物资源库和信息服务平台，扩大资源储备”。从这一要求来看，未来我国人类遗传资源的系统结构与管理模式优化是重要的基础性工作，是未来人类遗传资源深度利用与价值挖掘的必要前提。由于人类遗传资源信息库规模庞大、价值极高，因此应从情报学的角度提供有高度科学性的资源库管理方案，作为一种共享性质的资源平台与普通单纯信息资源平台还有明显的差异性，单条遗传资源信息包含了生物样本和相关信息资料[2]。这意味着资源使用或操作后可回溯性大大下降，任何以此错误的操作都可能造成这一重要资源的浪费或损失。

由此来看，在中国人类遗传资源平台样本信息的管理中有两项重要目标：一是考虑人类遗传样本信息资源的共享、便捷管理、信息安全保障需求，需要提供基础的资源标识符管理方法，解决样本管理与转移责任人的关联[3]，实现责任人精准定位；二是考虑人类遗传样本信息资源中生物资源部分的唯一性、不可回溯性特点（例如人类遗传样本中血液、人体组织等实物资源大多仅可供一次使用，此类资源无法重复利用），需要保证样本的所有变更操作、转移操作必须有准确的时间、操作人记录，提供有效的追溯功能[4]，以便于在使用时对责任人的追溯，同时实现对样本可用性保障、稀缺资源的节约利用保障。

人类遗传数据资源是重要的基础数据。当前我国人类遗传数据管理的重点集中在遗传资源基础数据的整合、标准体系建设、遗传资源数据库建设与共享机制的基础工作建设。遗传数据资源管理标准工作在数据存储规范、遗传信息描述规范和数据质量控制已经逐步形成了标准体系平台。但是遗传资源标准化数据库建设中数据安全、科学管理和高效共享在我国人类遗传资源库管理研究中还没有提出有效标准的研究方法。对于我国人类遗传数据资源的安全存储和共享使用，遗传资源在使用过程中的数据访问节点监控、数据可访问性保证、数据唯一性存储与共享都需要进行标准化的研究工作。

因此，结合以上两点需求，本文着重对我国人类遗传资源库管理现状、标识符技术研究与应用现状、情报学溯源管理技术研究现状等进行分析，综合探讨现有技术的特色优势与自有缺陷，专门针对中国人类遗传资源平台的管理需求，提出基于数字对象唯一标识符（Digital Object Unique Identifier，DOI）协议和开放源代码的版本控制系统（Subversion，SVN）版本溯源协议思想的统一资源标识符版本管理的创新方法。本文首先阐述人类遗传资源管理现状，介绍人类遗传资源管理的流程，然后提出统一资源标识符管理遗传数据资源的概念和方法，最后基于DOI资源标识符的方法，结合SVN和GIT方法提出改进的资源标识符方法，并对改进的标识符管理方法进行分析。

2　相关研究与应用

2.1　我国人类遗传资源管理现状

遗传资源信息管理中，国内主要是参考国际上的通用方法，侧重于数据的管理。管理数据方式采用的是传统的关系型数据库存储遗传数据关系与数据记录。遗传数据的使用主要是对存储于关系型数据库中资源样本进行分类浏览与检索。而对于遗传数据的演化、传播与历史溯源，国内外的研究都较少。

目前，国内在人类遗传资源管理研究主要集中在资源的共享与利用标准化上。其中涉及两个方面：一是资源样本的处理规程标准，包括人类遗传资源信息的采集技术规范（涉及标本采集、整理和保存）、信息加工技术规范（涉及标准的加工、备份）、资源库的架设规范等；二是信息的描述标准，包括共性信息的描述规范、特性信息的描述规范、个性化信息的描述规范等。

当前遗传数据管理的现状是基础存储与共享规范和平台已经初步完成，但是对于遗传资源

数据的高效分享是让遗传数据资源发挥重要作用的前提。将遗传数据资源高效共享给医学研究人员、研究人员基于庞大的遗传资源库进行数据分析和验证、产生相关医学研究成果是中国人类遗传数据资源管理平台建设的重要意义之一。在庞大遗传资源数据分享中，数据的安全性、数据的完整可访问性以及数据更新机制和数据管理责任划分都是需要提前解决的问题。本文基于中国人类遗传资源数据管理平台的基础工作流程，对上述需要解决的重要问题进行思考和分析，并提出一种可行的研究与解决方案。

在信息标准化管理下，信息采集、加工和处置的基本流程如图 1 所示。

从图 1 所示的流程管理模式来看，中国人类遗传资源平台能够分别针对生物样本的实物资源以及样本相关的遗传信息数据资源进行标准化处理，提供了在数据信息资源管理中共性信息、特性信息、个性化信息的基本规范[5]。其中，个性化信息规范能够为资源管理、操作、转移的责任人归属提供必要的管理条件。但从整个应用流程来看，其中并没有涉及资源操作和时间、责任人溯源规范，或者说在溯源方法上没有提供一个标准的方案。因此，本文的重点就在于通过优化资源标识符版本管理方法来实现样本信息的可溯源功能。

2.2　统一资源标识符管理基本方法

资源唯一标识符是一类统一资源标识符（URI）的统称，其目标是通过对数字对象的公认标准管理来永久性地标识某一个数字化对象。目前，在世界范围内应用最广、影响力最大的资源唯一标识符规则是DOI协议[6]。该技术协议早期主要用于对知识产权对象标准化记录，一方面实现知识产权的保护，另一方面也能够实现对相关数字资源的规范化管理。

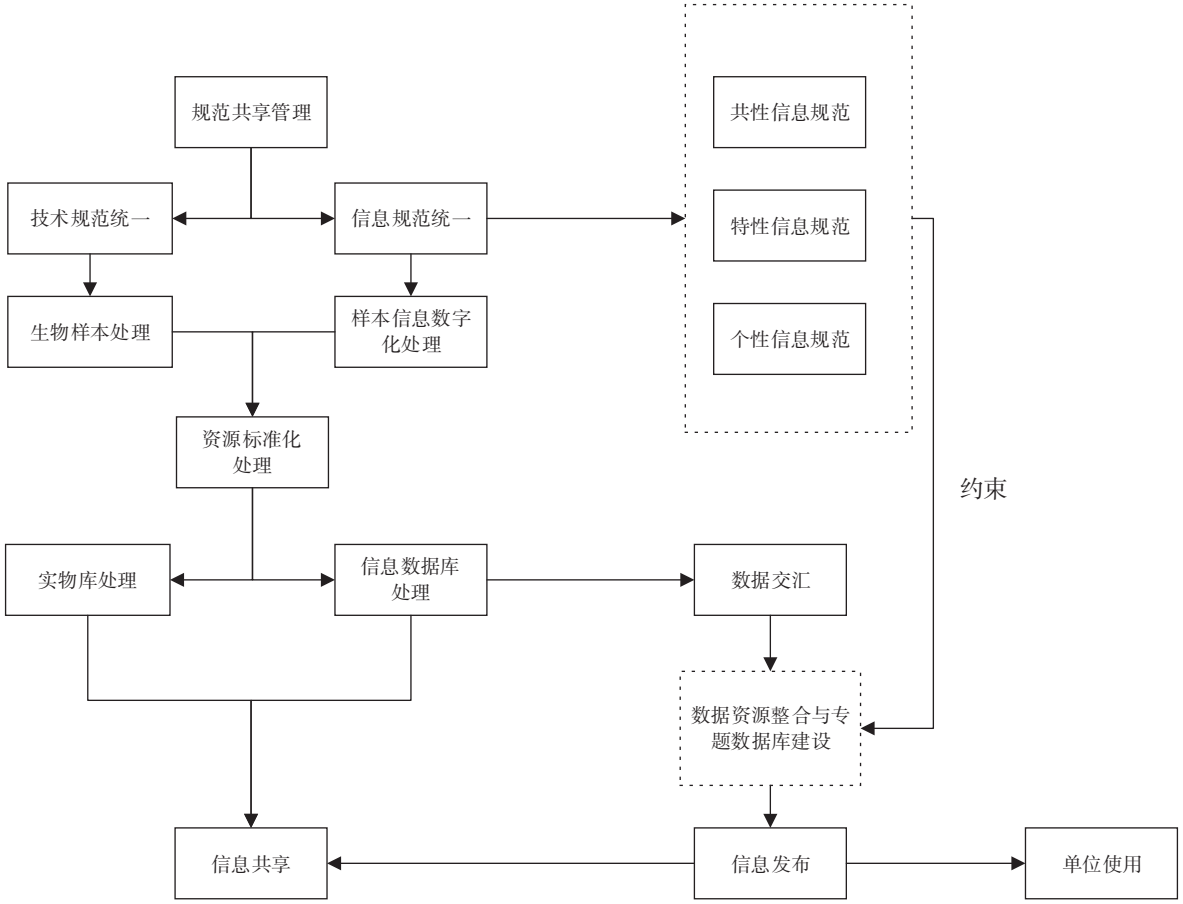

图 1　中国人类遗传资源平台资源共享的标准化管理基本模式

DOI的结构标准为“<DIR>.<REG>/<DSS>”。主要由3部分组成，即前缀、“/”分隔符、后缀。其中，前缀通过小圆点分割为两个部分：<DIR>为机构名，用于Handle标识系统（一种专门用于DOI标识识别的系统）的DOI命名机构（这里主要指中国人类遗传资源信息管理中心），固定值10；<REG>为注册机构代码，由DOI管理机构负责分配，一般为3～4位阿拉伯数字。后缀主要面向出版者，提供出版者的产品唯一性标识，规则不限。

以假定的0001a样本为例，DOI协议的标准标识符可以命名为“10.200/is NICGR 00−01−A”。其中，<DSS>为实现平台内部资源唯一标识的关键，“is NICGR00”为机构编号，后续分别为逐级分类编号。一般在该部分的规范设计中的难题是如何提升下级细节目录分类的精细化和识别唯一化，其设计思想与整个唯一标识符的规范思想相一致。目前，国内外对DOI模式的应用研究大都针对<DSS>字段的规范设计[7]、在线加入与查询机制实现[8]、中文化DOI应用策略等。

仅从本文研究来看，现有的DOI协议标准的优势在于便于实现大规模资源的统一资源标识符管理。相比于URN、SICI、BICI、PII、PURL等唯一标识符协议。该协议有两个特点：一是能够实现对持久性的唯一标识，这对于人类遗传资源这类有较高安全性和知识产权保障需求的资源信息管理有明显优势；二是现有的兼容性条件、互操作性条件十分显著，通过常见的Handle系统就能够实现简单的系统假设。

从DOI协议的特点来看，DOI协议非常适合中国人类遗传资源平台样本信息的管理，但是该协议也存在一个明显的缺陷问题，即无法提供可溯源功能。该协议早期的设计思想是针对版权和知识版权保护而提出的，资源本身的高可备份性和只读性决定了收录后的资源在使用中出现的修改、移动、删除等操作大多具备可回溯能力。但人类遗传资源则有其特殊性，收录后的操作、转移和使用必然会在可预见的时间内出现，这就导致资源流转过程中出现不同时间的差异，因此需要加入有效版本管理方案，这是现有DOI协议所无法完成的功能。本文旨在对传统DOI协议进行改进，提出一种可供溯源的<DSS>字段命名方案，实现这种有资源记录溯源需求的管理。

2.3　溯源管理基本方法与意义

溯源管理的基本思想是，区分并记录不同操作时间后的资源单独记录，即通过版本来区分标识。目前常见的版本管理模式主要分为两类：一是集中版本控制。这种模式下版本库存放于服务器，版本获取过程中需要从服务器读取数据[9]。类似于软件设计中普遍的浏览器端/服务端模式（Browser/Server，B/S）构架思想，常见的集中式版本协议主要为SVN协议。二是分布式版本控制。与前者相反，保证用户端或备用服务器都保存完整版本库[10]，即便服务器故障或者个别用户端故障都不会影响版本库的获取。这类似于软件设计客户端/服务端模式（Client/Server，C/S）构架和分布式数据库的设计思想，常见的分布式版本协议主要为分布式版本控制系统（GIT）协议。

单从两类协议对比来看：SVN协议与GIT协议均为开源协议[11]，实现成本都比较理想；SVN协议具备协同管理、旧版本还原、日志修改与查询、版本差异对比等功能[12−13]，目标功能需求的实现能力较强，但对网络的依赖度较高；GIT协议基于版本库树形分类来实现索引[14−16]，具备基本版本管理功能外还提供了强大的分支管理，分支管理的最大特色是减少对服务器的依赖，甚至能够实现不联网处理，但这种方法也意味着安全性的下降。综合来看，GIT协议的特色在于提升版本管理的稳定性，但不能保证版本管理的中央服务器授权，因此可以借鉴GIT协议的中央控制特色，结合具备集中控制的SVN对DOI进行版本管理。

DOI结合具备集中控制的SVN方法进行版本管理非常适合中国人类遗传资源平台样本信息的管理，同时也为人类遗传资源数据提供可溯源功能。在实际遗传资源数据管理工作中，通过对数据资源版本控制，可以对数据的所有迁移和更新历史进行追溯和定位。平台中的数据所有历史副本将被永久存储，相关的数据维护提交人、数据

访问用户都有对应的记录。如数据发生泄漏和错误，可以通过历史版本与溯源追溯到责任人，同时也可以恢复和找回历史数据资源。通过这一中央控制的溯源机制，使数据安全存储得到保障，数据资源的共享与传播控制效率提高，数据管理的责任边界可以清晰定位。

3　可溯源管理方法探析

结合前文分析来看，目前中国人类遗传资源平台样本信息管理的关键性目标在于实现便捷搜索和可溯源功能。在便捷搜索方面，最优的标识符方案为DOI协议方案，但该方案本身不具备版本管理能力，无法提供可溯源功能。因此，需要设计独立的<DSS>字段命名规范，提供内建的版本识别规范。在可溯源功能方面，在基本的字段命名方案基础上，需要对版本协议进行优选。考虑到平台信息共享与服务的稳定性以及人类遗传资源信息的保密性，具体的版本管理协议可以采用具备中央控制能力SVN。

3.1　基于DOI资源标识符的可溯源管理优化方案

结合可溯源目标的版本管理需求，在原有的DOI资源标识符中加入版本区分代码。这里主要参考国家卫生和计划生育委员会（简称“卫计委”）组织分类标准进行设计。表1为标识符方案优化对比说明。

在该标识符设计思想下，能够对版本信息进行标识，同时对版本信息和资源基本信息进行分离。在这种情况下不影响Handle系统的资源识别，只需要专门对<DSS>的识别规则进行定义即可满足正常应用需求，同时可以提升信息溯源的有效条件。

3.2　基于具有集中管理功能的SVN版本管理方案的可溯源管理实现方法

版本管理的基本思想是基于GIT分布式版本信息管理与SVN协议权限控制功能的融合。

首先，在Handle系统的基础上进行SVN版本管理，对于资源共享平台管理单位、申请单位、保存单位、使用单位人员进行不同的权限管理。具体需要在SVN服务器段进行权限配置，需要根据DOI的版本号映射成SVN的版本号，需要实现SVN协议和DOI协议的桥接。

其次，在子服务器端创建SVN版本库。服务端版本库管理借鉴GIT协议的集中管理思想，但同时需要剔除分布式管理功能，即：服务端不接受其他终端或资源保存单位的版本库，终端信息记录后服务器仅记录增加版本，不提供任何版本信息的删除和其他修改功能。在服务器新增记录更新后其他需要保存版本信息的子服务器通过“svn_doi reset”命令将获取新资源标识符并转换为SVN版本信息[17−18]，用于更新本地版本库，依据所获取的版本信息来确定新版本信息位于版本树形结构的归属于哪个索引头和内容之下[19]。总的来说，就是借鉴了GIT协议的版本集中管理思想，但需剔除分布式同步更新功能（避免无中央服务器操作权限的使用者通过修改子版本服务器信息来篡改中央服务器SVN版本库，从而保证资源使用记录的安全性），子服务器的版本树与记录更新则利用SVN思想进行更新，保证分支机构在资源信息查询时能够获得所需的完整版本信息。版本管理流程图见图2。

表1　DOI资源标识符优化方案对比

方案名称	DSS字段标准	0001a 2017年12月17日操作（总第n次操作）后样本实例	特色说明
DOI标准方案	<DOI命名机构>.<注册机构代码>/<自定义资源描述>	10.200/is NICGR 00−01−A	可实现基本的资源唯一标识
卫计委组织分类方案	<DOI命名机构>.<注册机构代码>/<自定义资源描述><机构><分类><时间戳><流水号>	10.200/is NICGR 00− 01− A−20161217−n	能够在基本的资源标志外实现时间操作分类
改进后的NICGR资源样本命名方案	<DOI命名机构>.<注册机构代码>/<自定义资源描述><机构><分类><时间戳>/<版本号>	10.200/is NICGR 00−01−A−20161217/rn	能够实现明细化的版本分类，提供多版本信息回溯的基本条件

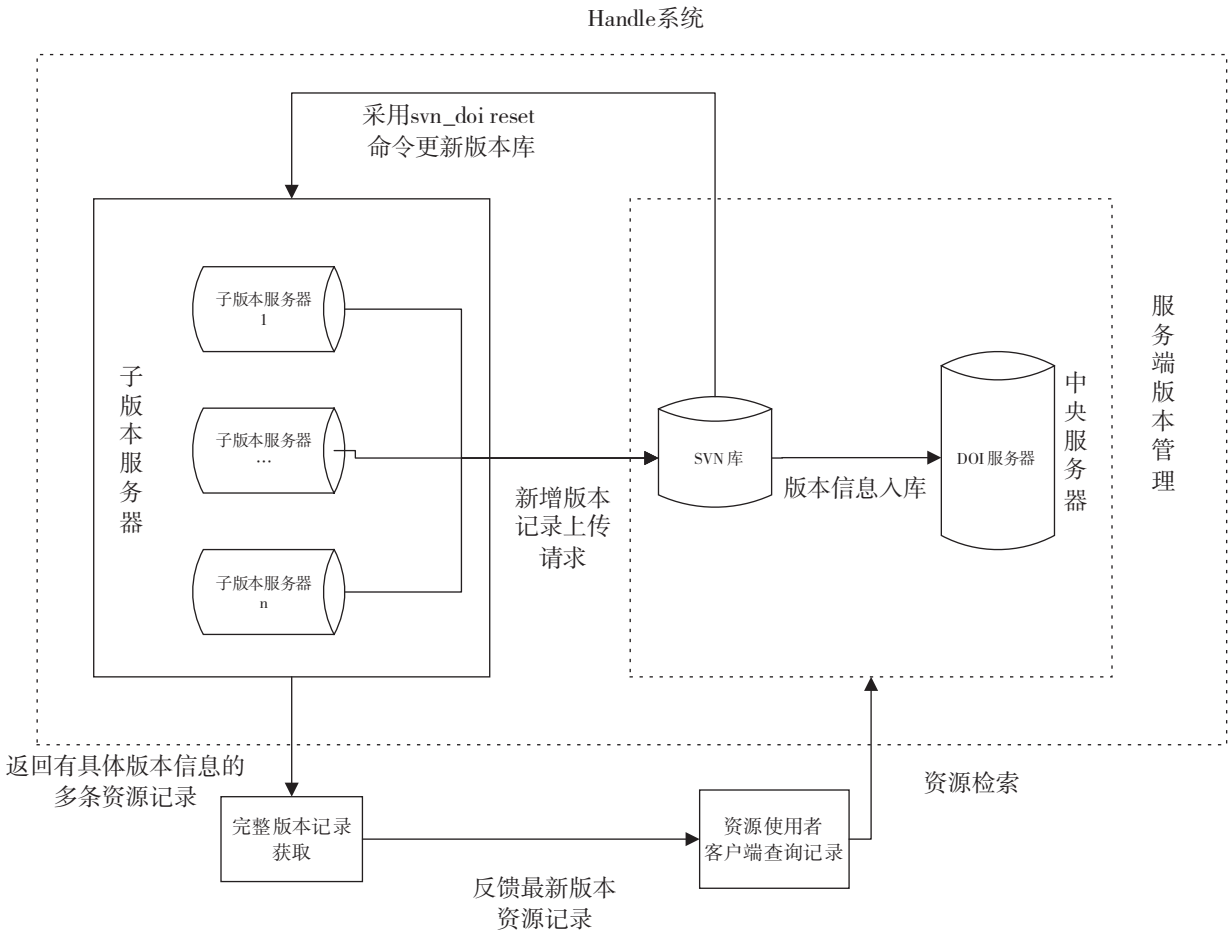

图 2　资源标识符版本管理与获取流程示意图

4　总结与展望

本文结合中国人类遗传资源平台样本资源标识符管理的需求进行分析，探讨了平台资源管理的发展现状和问题，提出了改善资源版本标识的新需求。结合唯一标识符技术、版本管理技术的理论研究与应用现状，指出在目标应用场景下所存在的问题，提出了重新设置DOI协议<DSS>字段命名规则中加入版本号的具体方案，最后对DOI+SVN模式下版本管理的基本思路进行说明，提出了具备中央控制能力的SVN协议权限控制思想的版本管理模式。通过对数据资源版本控制，可以对数据的所有迁移和更新历史进行追溯和定位。这一中央控制的溯源机制可以保障数据安全存储，提高数据资源的共享与传播控制效率，清晰定位数据管理的责任边界。

通过资源标识符的版本记录功能可以实现对人类遗传资源信息的细化记录：一方面能够保证资源操作记录的完整性，满足资源样本多样化利用记录的可行性；另一方面可以保证所有的资源操作记录都可被记录，稀缺资源的不合理使用情况均可被查询，在后期的资源使用追溯中能够更加准确地定位责任人，不会出现样本资源浪费却无据可查的问题。

进一步研究遗传资源数据可溯源方法和完善实施方案，为遗传资源数据的保护和高效共享提出了一种安全有效的实施办法。DOI资源标识应用和SVN中央控制的版本控制机制的结合创造性地为遗传资源数据的定位、溯源与管理提供可靠的方法流程。

在未来的目标用户资源查询中可以采取如下

（下转第 36 页）

[3] 国务院办公厅关于建设第二批大众创业万众创新示范基地的实施意见[EB/OL].(2015−06−21)[2017−06−21].http: //www.gov.cn/zhengce/content/2017−06/21/content_5204264.htm.
[4] 中国创业创新指数（2015−06−21）[2017−06−21].http://inno.36kr.com/demo.html#/province_rank.
[5] 扬州“万方科创书院”正式开业[EB/OL].(2017−05−19)[2017−06−02]. http: //www.yangzhou.gov.cn/bmdt/201705/RESJSBPZ94U3G53O9J3GC1P0B1H−PFV54.shtml.
[6] 书院[EB/OL].(2017−06−02)[2017−06−02].http://baike.sogou.com/v217267.htm.
[7] 张芳, 吴颖利. 论中国古代书院与高校图书馆[J]. 唐山学院学报, 2010, 23(4): 107−108.
[8] 马国柱. 关于文化产业发展路径的思考[J]. 中国出版, 2012(19): 24−28.
[9] 刘援朝, 刘景. 浅论我国文化产业的发展及对策[J]. 社会科学论坛, 2003(6): 73−75.
[10] 王志宏, 杨震. 人工智能技术研究及未来智能化信息服务体系的思考[J]. 电信科学, 2017, 33(5): 1−11.
[11] “万方科创书院”启动运营: 探索“科技文化融合发展城市综合体”模式[EB/OL].(2017−05−22)[2017−06−02].http: //www.ce.cn/cysc/tech/gd2012/201705/22/t20170522_23094780.shtml.

（上接第 24 页）

流程：第一步，用户确定资源信息，自行判断是否已获知版本信息；第二步，如果获知版本信息则直接输入标准DOI识别符号并返回相应版本文件，如果未获知版本信息则返回所有版本文件。所有返回文件按照标识符版本编号标准进行额外命名（示例：资源信息名称rn.pdf）。

中国人类遗传资源平台在未来发展过程中将会有更大的资源积累和更多的新资源出现，在开放化资源共享的发展趋势下，未来样本信息管理工作需要进一步完善，尤其要注重对资源信息的查找效率、查找精确度、责任版本记录等关键信息，保证资源共享服务的综合质量和已有资源的安全保障质量。在人类遗传资源与生物信息领域样本资源存储国际标准竞争中，随着我国存储标准逐步完善、管理数据规模化、管理经验成熟化，基于DOI的样本信息可溯源的标识方法将产生重要的影响力，成为国际遗传信息资源标识的标准。本研究也希望能够为平台样本信息资源管理工作的改良提供一定借鉴。

参考文献

[1] 国务院. “十三五”国家科技创新规划[A].2016.
[2] 尹姗姗,李向旭,聂伟.河南厅直医疗卫生单位人类遗传资源存储管理现状分析[J].河南医学研究, 2015(2): 79−82.
[3] 肖红.国内外数字资源唯一标识符系统对比研究[J].科技情报开发与经济,2016,1(6):140−143.
[4] 杨亚军.人类遗传资源信息的数字化关键技术[J].中国科技成果,2015(5):11−12.
[5] 王帅.基于DCI的版权保护核心架构研究与实现[D].北京:北方工业大学,2015.
[6] 肖红.国内外数字资源唯一标识符系统对比研究[J].科技情报开发与经济,2016(6):140−143.
[7] 邢瑞华,杜一诗,张晶.信息与文献标识符标准的推广与应用[J].出版参考,2014(28):33−34.
[8] 刘丹军,付鸿鹄,文奕,等.科技知识组织体系版本管理系统设计与实践应用[J].现代图书情报技术, 2015, 31(4):79−86.
[9] 刘校妍,蒋晓敏,楼燕敏,等.基于事件和版本管理的逆基态修正模型[J].浙江大学学报理学版, 2014, 41(4):481−488.
[10] 冯嘉俊,赵海燕.分布式版本管理工具:Mercurial[J].数字技术与应用,2015(1):224−225.
[11] 曹阳,黄海峰,梁成辉,等.CIM/E 模式版本管理和映射方法[J].电力系统自动化,2015(8):149−154.
[12] 许磊,夏翠娟,刘炜,等.关联数据URI设计规范探讨[J].国家图书馆学刊,2016(5):22−32.
[13] 刘啸.主流源码版本管理工具的特色浅析[J].程序员,2008(3):116−118.
[14] 刘乐.软件项目管理与SVN[J].科技信息,2011(21):83.
[15] 郑重华,苏振涛,陈磊,等.基于SVN的通信机房图纸管理方法研究[J].电信技术,2016(4):52−55.
[16] 汪玉.个性化云同步系统中版本库副本创建和选用策略及模型研究[D].合肥:安徽大学,2016.
[17] 蒋明浪,程方.基于RESTful的泛在网资源标识防篡改策略[J].电视技术,2014,38(19):69−71.
[18] 戴宇.针对REST架构的WebService研究[J].无线互联科技,2016(15):113−114.
[19] 曲云鹏.存档资源键研究[J].数字图书馆论坛, 2014(12):29−35.