Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于唯一标识符的多类型信息资源共享系统构建*
——对国家自然科技资源e-平台建设方案的思考

□ 彭洁 / 中国科学技术信息研究所 北京 100038
/ 武汉大学信息管理系 武汉 430072
赵辉 王运红 / 中国科学技术信息研究所 北京 100038

摘要：标识符体系是信息管理系统中常用的一种技术体系，对资源描述和发现具有重要意义。文章在梳理国际常用的标识符系统的基础上，针对自然科技资源的特点，制定了国家自然科技资源e-平台的标识符体系的方案，并对这种方案的应用效果进行了评述，提出了改进意见和建议。

关键词：唯一标识符，信息资源共享系统，国家自然科技资源e-平台
DOI：10.3772/j.issn.1673—2286.2013.07.007

1 引言

在科学实验和研究活动中会用到实验试剂、动植物种质、微生物菌种、生物标本、岩矿化石标本、标准物质等，这些被定义为自然科技资源[1]。自然科技资源的应用价值很高，多由专业机构进行集中收集和保存。一直以来，自然科技资源的收集和保存情况基本通过收藏机构出版的动植物志或目录等方式公开。互联网普及以后，建立具有检索功能的自然科技资源网站也成为主要的信息公开渠道。但是，由于自然科技资源的种类丰富，保藏机构繁多，迄今为止，我国还未建立起各类自然科技资源统一检索的网站为科技工作者提供服务，因此，在国家科技基础条件平台计划中，启动了国家自然科技资源e-平台的项目，旨在把分散在全国各地的八大类自然科技资源信息进行整合，提供集成检索服务。这项工作开展的关键基础性工作之一就是构建自然科技资源唯一标识符体系。

2 标识符体系的基本概念和原理

在数字环境中，标识符是在信息分类的基础上，将信息对象（编码对象）赋予具有一定规律的、易于计算机和人识别处理的符号，以便正确地定位和管理该对象[2]，有时被称为“客体标识符”[3]。这种标识符要求在一个信息系统中是唯一的，国际上对信息资源标识方案中，使用较为广泛的有数字对象标识体系（Digital Object Identifier，DOI）和统一资源标识符（Uniform Resource Identifier，URI）等。资源标识对资源的描述和发现具有重要意义，用以实现资源在网络环境下的唯一识别[4]。

一般来说，标识符系统应该包括名称空间、唯一标识符、命名机构、命名登记系统和解析系统5个部分[5]。

2.1 统一资源标识符

统一资源标识符（Uniform Resource Identifier，URI）[5]是一个互联网标准，属于请求注解（Request for Comments，RFC）档案式文档系列。所有的URI都要在互联网地址编码分配机构（Internet Assigned Numbers Authority，IANA）注册。URI规定了统一资源标识符的语法构成。

在URI体系下，有两种主要的应用，分别是统一资源定位符（Uniform Resource Locator，URL）和统一资源名（Uniform Resource Name，URN）。三者的关系如下：

* 本文受“中央级公益性科研院所基本科研业务费专项资金”（编号：ZD2012-6-1）支持。

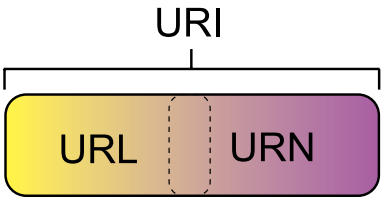

图1 URI、URL和URN的关系（引自维基百科）[6]

URN是一种与资源位置无关的标识符，其语法规则使其他命名空间的资源能非常简便地映射到URN空间中。URL则用于在互联网上指示资源所在的地址。URN和URL的联合使用，可以提供资源的识别与定位服务。URI统一解析的基本思想是实现一个解析发现系统（Resolution Discovery System或Resolver Discovery Service，RDS），使之工作在各种标识符解析系统之上，为各种标识符提供统一的解析入口。当用户向RDS提交一个URN后，有RDS负责发现解析该URN的实际解析系统，并按照合适的规则将解析请求发送给该系统，之后真正的解析系统将标识符解析后返回给客户。但是URI统一解析的实现还在研究与试验过程中，还不能实际进行使用。在实际应用中，URI被许多系统分别用来进行资源标识和定位，并使用URL重定向机制解决URL更新问题。

2.2 数字对象唯一标识符

数字对象唯一标识符（DOI）是由美国出版协会（AAP）提出并建立的标识系统，是为了解决数字对象的跨系统操作，针对网络环境中数字对象具有移动、易变性和多源性的特点，用一种持久稳定的数字对象定位方案来保证数据对象的地址长期可访问[7,8]。现在DOI由国际DOI联盟（IDF）统一管理和维护，各个地区的DOI注册机构（RA）在IDF的管理下，开展DOI命名的分配、解析、描述和管理。

与URI相比，DOI不再仅仅是一种技术标准，其应用和推广严格遵循了唯一标识符的开发和使用要包括名称空间、唯一标识符、命名机构、命名登记系统和解析系统5个部分的要求。主要表现在：

（1）在名称空间上，DOI已经正式成为ISO标准体系的一部分。

（2）在组织上形成了三级结构，最上层是IDF，负责全球DOI的管理和统一解析服务，向各注册机构（RA）分配DOI前缀；第二层是一批RA。RA负责某一区域或行业的DOI注册和服务，管理一批DOI的前缀。第三级是出版代理（PA）。PA负责生成每个DOI，维护DOI的可用性维护等。

（3）开发了DOI解析系统，提供注册程序接口和工具。

因此，DOI的推广和使用比URI更具有全球统一性和可用性。

2.3 已有唯一标识符体系在自然科技资源e-平台建设中遇到的挑战

从上述资料可以看出，在实际应用中，DOI体系比URI具有更强的可用性。考虑到自然科技资源的战略性和国家安全的需求，国家自然科技资源e-平台（以下简称“e-平台”）可以借鉴DOI体系在我国国内建立起自然科技资源在互联网环境中的唯一标识体系。但还需要解决自然科技资源实物资源与网络信息目录、多媒体展示图片、音视频文件的相互对应，以及一种自然科技资源在多个机构保存的对应规则问题。具体来说，即：

（1）每个自然科技资源保存机构在资源数字化过程中，会制作资源目录、资源图片、资源视频、资源音频等文件，在资源信息服务系统中，目录信息、图片信息、视频信息、音频信息对资源用户来说，具有不同的用途，因此既需要将目录信息与图片信息、视频信息、音频信息建立关联，又要进行分别管理。

（2）每种资源可能被多个机构保存，这些不同机构保存的资源，既可以认为是由不同的机构进行采集、加工和制作的不同版本、不同质量的不同资源，被分别赋予毫不相关的DOI号，也可以认为是同一种资源的多个备份，被赋予有关联关系的一组DOI号。

以上问题需要在自然科技资源e-平台建设过程中予以解决。

3 国家自然科技资源e-平台唯一标识体系方案及思考

3.1 e-平台唯一标识的对象及特点

自然科技资源共享是一个复杂的系统工程，涉及资源单位、资源用户、政府部门、国家、社会等多个主体，发生整合、利用、监督、服务等多种行为，关系个人利益、国家利益和社会利益等多元利益[9]。国家自然科技资源e-平台整合了植物种质、动物种质、微生物菌种、人类遗传、生物标本、岩矿化石标本、实验材料与标准物质共八大类自然科技资源。

八大类资源完全不同，自然科技资源的多元化决定了在对其进行唯一标识时，需要充分考虑每类资源的特殊情况。自然科技资源数据是八个数据库，数据类型既有结构化的共性描述规范数据，有的资源还有图片、音频、视频、动画和3D模型数据。共享描述规范是自然科技资源共享平台中对自然科技资源共性及身份描述的统一标准，包括各类资源共享的护照信息、标记信息、基本特征特性描述信息、其他描述信息、收藏单位信息和共享信息[10]。

下面详细介绍自然科技资源的八大类资源特点以及在e-平台上存储的数据情况。

（1）植物种质：植物种质是所有携带遗传物质的活体，不仅包括种子，还包括植株、根、茎、胚芽和细胞等等，甚至是DNA片段[11]。在e-平台的数据库中，每份植物种质资源的信息有结构化的共性描述数据，部分资源还有对应的图片，在虚拟博物馆中，还有部分植物种质资源的视频数据。

（2）动物种质：动物种质资源既包括已知的种质或者遗传物质，也包括一些遗传潜力材料。在e-平台上共有寄生虫、经济昆虫、水生动物、特种经济动物、畜禽五大类动物种质资源的信息。因为动物种质资源信息的特殊性，在e-平台上，既有动物种质结构化的共享描述数据，又有图片和3D模型数据，在虚拟博物馆中，还有部分植物种质资源的视频数据。

（3）微生物菌种：参与e-平台的微生物种质资源保藏单位，都已建有微生物菌种保藏中心，微生物菌种数据与实物菌种一一对应，主要是微生物菌种共性描述并发数据和部分菌种的图片数据。

（4）人类遗传资源：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[12]。在e-平台上共享的人类遗传资源主要是人体物质资源、重大疾病资源和少数民族资源，主要是共性描述信息和图片信息。

（5）生物标本：e-平台共享的生物标本信息有菌物标本、动物标本和植物标本，主要是共性描述信息和图片信息，在虚拟博物馆中，还有部分的视频数据、3D模型数据。

（6）岩矿化石标本：岩矿化石标本包括化石标本、矿石标本、岩石标本和矿物标本，因为岩石晶体的化学结构非常复杂，所以在数据存储时考虑化石拉丁名、矿物晶体结构等专业描述符号的表达与存储，图片更多。岩矿化石标本除了必须有共性描述信息外，图片数据量非常大，还有部分标本的视频数据和3D模型数据。

（7）实验材料：从应用和共享服务的角度出发，依据不同实验材料所具有的特殊属性和应用属性进行分类，同时根据目前的工作基础和科技发展对实验材料的需求，e-平台上的实验材料资源仅限于实验动物、实验细胞和微生物培养基。在e-平台上的数据有共性描述规范数据，实验动物资源有较多的图片数据。

（8）标准物质：按照国际标准化组织指南30和国际通用计量学基本术语定义，标准物质（Reference Material，RM）是具有一种或多种足够均匀和很好确定的特性值的、用以校准设备、评价测量方法，或给材料赋值的材料或物质。这决定了标准物质的特殊性，既可以物质形式存在，也可以是计量方法等，数据存储主要是结构化的共性描述数据。

从上述情况可知，自然科技资源种类繁多，数量巨大，具有丰富性和独特性，承载着人类对自然资源的科学认识或技术评价。自然科技资源收集和保藏的机构很多，各机构之间的保藏资源有重复，但是也有很多共享和业务联系，比如，植物种质和动物种质的培育，标本的互借等。因此唯一标识符的建立最好既能体现出不同机构收藏的资源具有唯一性，也要方便展现出不同机构间收藏的重复性资源，以及不同类资源的相关性。因此，e-平台资源的唯一标识符的命名规则确立了以“资源分类码+保藏机构代码+资源顺序码”的编码方案，其解析规则定为以唯一标识符为核心，辅之以资源分类和关键词来展示资源的相关性。

3.2 e-平台资源唯一标识符命名规则

e-平台资源实现属地化管理，即由资源拥有单位或个人对所拥有的资源及其资源信息进行全权管理。因此，平台资源的唯一标识符的命名规则是：

资源分类编号（2位）+单位所在区域编号（2位）+资源单位性质代码（P或C）+资源保藏单位/人序号（4位）

各部分的编码规则如下：

◆ 资源分类编号

植物种质资源：11	动物种质资源：13	微生物菌种资源：15	人类遗传资源：17
生物标本资源：21	岩矿化石资源：23	实验材料资源：31	标准物质资源：33

◆ 单位所在区域编号

代码	省市名称	代码	省市名称
11	北京市	44	广东省
12	天津市	45	广西壮族自治区
13	河北省	46	海南省
14	山西省	50	重庆市
15	内蒙古自治区	51	四川省
21	辽宁省	52	贵州省
22	吉林省	53	云南省
23	黑龙江省	54	西藏自治区
31	上海市	61	陕西省
32	江苏省	62	甘肃省
33	浙江省	63	青海省
34	安徽省	64	宁夏回族自治区
35	福建省	65	新疆维吾尔自治区
36	江西省	71	台湾省
37	山东省	81	香港特别行政区
41	河南省	82	澳门特别行政区
42	湖北省	99	不详
43	湖南省		

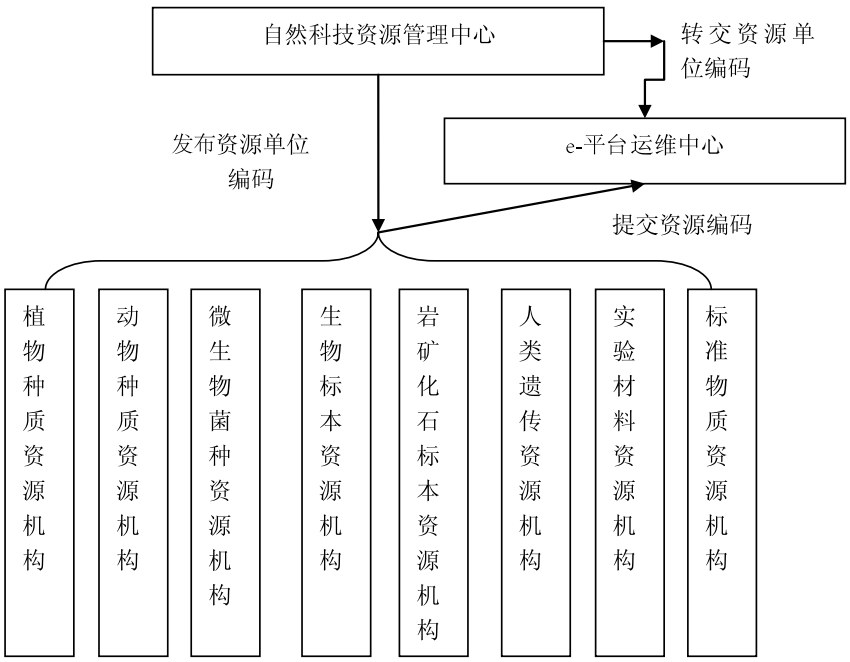

图2 e-平台的资源命名机构

◆ 资源单位性质代码

P（Person）——表示资源提供者是自然人；

C（Corporation）——表示资源提供者是法人实体。

◆ 资源保藏单位/人序号

资源保藏单位/人序号排名不分先后，只依据数据进入e-平台数据库的先后顺序来编号，具体序号由国家自然科技资源平台管理联合办公室统一给出。

例如，中国农科院品资所保藏的植物种质资源编号为：

1111C0001000000001，1111C0001000000002，……

e-平台方案的优缺点分析：

e-平台现有的编码方案的优点在于，明晰了各类自然科技资源信息的所有者和责任单位，管理和维护职能方便。其缺点在于：e-平台现有的编码方案仅能区别出一级资源类及其保藏机构，而不能区别出四级以下资源在哪里保藏。如果在编码方案中置入四级分类编码，则可实现四级分类资源的标识、指示和解析服务。

3.3 命名机构及其管理

e-平台的资源命名机构采用三级管理的方式。各级机构的关系如图2所示。

第一级，是在全国设立一个资源机构管理中心，负责对全国各地资源保藏机构进行认证，为通过认证的资源机构编制单位编码，并把确定的唯一标识符前缀信息告知e-平台运维单位，负责组织对各类资源的质量审查工作。该中心由科技部农村中心承建并运行[13]。

第二级，是e-平台运维中心，负责管理和运行e-平台，收集各资源单位提交的资源信息文件，上载到信息管理系统中，并进行发布。

第三级，是各资源保藏机构。这些机构负责将自己保藏的资源按照e-平台资源规范进行分类，制作资源题录信息、图片信息、视频信息，并按照统一的格式要求赋予资源唯一标识码后，将这些信息定期编制成资源文件，提交到e-平台的运维中心，由运维中心负责将格式正确的资源信息文件正式载入到管理信息系统中。

e-平台对命名机构的管理满足了自然科技资源管理的需求。

3.4 登记系统设计

资源登记系统分为两级，一级是各资源单位将制作完成的资源信息文件，提交给e-平台运维中心，运维中心根据《国家自然科技资源数据上报规范》，编制了数据规范审查软件，对各资源单位提交的数据中的必填项、共性字段、个性字段的数据规范性、书写格式、内容格式、字段值的赋值类型、存在性及唯一性等是否符合规定等进行审查，以确保资源信息的基本可用性。

第二级是专家评价。经过格式审查后的数据，根据随机抽取规则抽取部分数据，随后进入专家评价系统，由各行业专家对数据内容的科学性和正确性进行审查，从而有效地保证了数据的科学、客观和真实可靠。

经专家审查通过的资源数据，才正式完成了资源的登记工作。

登记系统保证了e-平台数据的质量，有利于e-平台数据的管理和维护。

3.5 解析系统设计

国家自然科技资源e-平台基于唯一标识符的解析主要考虑三方面的功能。

1）根据关键词进行解析

主要针对资源名称、资源描述、资源关键词等字段提供检索结果。通过关键词的检索，可以得到资源名称、资源描述和资源关键词的检索结果。

2）根据资源分类进行解析

e-平台可以根据资源一级分类码，再辅之以关键词的检索，即可获得想要的资源信息。

3）多元信息的展示

在检索到想要的资源之后，系统除了提供资源题录信息外，还提供资源的图片信息、音视频信息、三维图像信息等，以便用户能多方位地获得想要的各类信息。

在e-平台的规则空间中，有单位编码和一级资源码的限制，使得e-平台的解析系统只能解析到一级资源，若使e-平台能解析到四级分类码，则做到四级资源的多项解析，从而为用户提供更加细致、全面的资源服务功能。

4 总结与展望

通过对e-平台唯一标识体系的梳理，可以看出e-平台已经初步完成了“有哪些资源分别在哪里”设计目标，但在唯一标识体系的管理和应用上，还有以下几方面工作需要深入研究：

（1）在资源编码的粒度上，目前的资源编码是以自然科技资源的一级分类+资源单位+资源的单位序号为基础的，这样的编码设置，不利于检查资源提供机构是否为每个资源提供了详细的资源信息。随着资源信息精细化加工工作的深入，e-平台导航系统还可以改进资源编码规则，细化编码前部的资源分类码到四级类，这样将有利于资源导航至更加细化的资源信息页面。改进方案就是将资源的四级编码嵌套到唯一标识码中，即可揭示同一种资源在不同机构的保藏情况。在此基础上，可以提供在多项解析基础上的关联信息展示等服务。

（2）采用DOI编码方案的e-平台唯一标识符是不定长编码，在互联网上传输的过程中，容易造成因传输原因导致的编码错误又无法校验的情况，为了保证用户检索信息获取的准确性，在未来的系统升级改造时，可以考虑这样的传输校验机制：连续发送三次编码，在用户端进行比对，取两次以上一致的结果为准，三次都不一致的将报错。

（3）在DOI编码确定后，可以在未来的e-平台服务中，增加面向科技论文参考文献的引用的服务内容。可以考虑参考文献引用的格式，采用“作者，资源信息发布机构，资源链接，DOI编码”的格式，方便用户对自然科技资源信息的使用。

参考文献

[1] 杜占元,刘旭,等.自然科技资源共享平台建设的理论与时间[M].北京:科学出版社,2007:4.
[2] 徐枫,宦茂盛.政务信息资源目录体系技术概述[J].信息技术与标准化,2005(11):23-27.
[3] 徐冬梅,吴东亚,姚忠邦.抽象语法技法与客体标识符介绍[J].信息技术与标准化,2006(1):44-45.
[4] 资源唯一标识符规范,基础科学数据共享网项目标准[EB/OL]. [2013-01-02]. http://www.nsdc.cn/upload/110526/1105261310257020.pdf.
[5] 毛军,孟连生,等.试论我国数字资源唯一标识符发展战略[J].现代图书情报技术,2005(2):1-4.
[6] URI scheme,维基百科[EB/OL]. [2013-01-02]. http://en.wikipedia.org/wiki/URI_scheme.
[7] PASKIN N. Digital Object Identifiers for Scientific Data [J]. Data Science Journal, 2005(4): 12-20.
[8] PASKIN N. Components of DRM Systems: Identification and Metadata [C]// Lecture Notes in Computer Science, Berlin: Springer, 2003.
[8] 贺德方,张旭.服务于科技信息资源共享的数字对象唯一标识应用研究[J].图书情报工作,2007(8):26-29.
[9] 王运红,张莞,沈欣媛.国家自然科技资源e-平台建设实践[J].中国科技资源导刊,2008(7):16-19.
[10] 曹一化,刘旭,等.自然科技资源共性描述规范[M].北京:中国科学技术出版社,2006:1.
[11] 卢新雄.植物种质资源库的设计与建设要求[J].植物学通报,2006(1):120.
[12] 国务院办公厅文件 国办发[1998]36号.人类遗传资源管理暂行办法(中英文).1998-06-10.
[13] 卢兵友.自然科技资源平台共享机制建设思考[J].中国科技资源导刊,2008,40(4):6-10.

作者简介

彭洁（1962-），研究方向：信息资源管理、科技资源管理。E-mail: pengj@istic.ac.cn

Multi-type Information Resources Sharing System Construction Based on the Unique Identifier - Rethinking National Natural Science and Technology Resources e-platform Construction Scheme

Peng Jie / Institute of Scientific and Technical Information of China, Beijing, 100038
/ Wuhan University, Wuhan, 430072
Zhao Hui, Wang Yunhong / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Uniform identifier system is commonly used in information management system, which is important for information resource description and finding is. After sorting out international common identifier system, this paper formulated the uniform identifier system schema of National Natural Science and Technology Resources e-Platform according to the characteristics of natural science and technology resources. Then the application effect of the scheme was described. Finally, a few improvement comments were put forward.

Keywords: Uniform identifier, Information resources sharing system, National Natural Science and Technology Resources e-platform

（收稿日期：2013-01-29）