# 数字图书馆中科学数据目录体系建设方案探讨

□ 涂勇 / 中国科学技术信息研究所 北京 100038
彭洁 / 中国科学技术信息研究所 北京 100038
/ 武汉大学信息资源管理学院 武汉 430072
郭晓峰 / 北京万方数据技术研究院DOI运行中心 北京 100038

摘要：文章对数字图书馆科学数据目录体系建设的必要性进行了分析，在国内外科学数据目录体系的发展现状对比研究的基础上提出了科学数据目录体系的概念，并从标准规范、技术架构和管理机制三个方面对数字图书馆中科学数据的目录体系框架进行了阐述。

关键词：数字图书馆，科学数据，目录体系
DOI：10.3772/j.issn.1673—2286.2012.11.011

科学数据指人类认识自然、利用和改造自然的各类科技活动所产生的基本科学技术数据和按照不同需求而系统加工的数据分析产品和相关信息[1]，科学数据是一种重要的创新资源，科技创新依赖于对科学数据的发现、提炼、归纳和分析研究。随着大数据时代的到来，科学数据作为一种珍贵的科研资产，是科技信息资源的重要组成部分，处于转型期的图书情报机构纷纷将科学数据作为重要的研究对象，良好的目录体系有利于对数据资产进行管理，便于用户进行检索和查找，其价值越来越被政府和公众所认识。

在数字图书馆中集成科学数据目录资源是科学数据资源管理方式的一种创新，数字图书馆作为一种公益性、中立的第三方载体，在尊重科学数据资源知识产权的前提下，能够很好地解决科学数据机构之间的利益冲突的问题，通过构建科学数据核心元数据，汇集最核心、最简单的科学数据目录信息，与其他类型科技信息资源进行有效的关联和引用，从而能更好地实现科学数据资源导航，发现并获取其他不同类型的信息资源，这是数字图书馆中建立科学数据目录体系的出发点。

## 1 数字图书馆中科学数据目录体系建设的必要性分析

### 1.1 科学数据资源的管理缺乏一个中立、综合的信息载体

专业性的科学数据资源网站由于领域、机构、功能定位上的限制，大量的科学数据资源散落在各科学数据机构之间，科学数据资源的获取花费了大量的人力、物力资源，其获取成本比较高，且部分科学数据实体资源还带有保密性质，同时考虑到学科分类的精细化以及机构利益的因素，且因此目前科学数据目录的提供方式主要是由专业科研院所为本学科领域提供纵深的科学数据共享服务，提供关于一个学科、一个领域的科学数据资源，由于在系统架构、标准规范方面的差异，无法与其他类型的信息资源进行有效的集成和整合。

科学数据综合性、导航性的网站由于定位和数据源限制，很难形成统一开放的科学数据目录体系。中国科学院科学数据库将整合中科院各专业科研院所的科学数据资源，形成了覆盖基础科学数据研究的科学数据目录体系，同时提供部分科学数据实体资源的存储、保藏和使用；国家自然科技资源平台集成了八大类的自然科技资源，提供资源的统一描述和导航。

虽然有专门性的科学数据目录体系，整个科学数据领域缺乏一个集中、中立的信息载体，提供相对集中的

不同种类科学数据资源的联合查询。数字图书馆为科学数据的管理提供了一种有效的载体。在网络时代，无论是文献还是数据都可以转化为数字形式，成为数字图书馆的管理对象并可在全球范围内传播和共享。

### 1.2 新型数字图书馆可与科学数据中心合作提供科学数据目录服务

数字图书馆作为数字资源保存、共享和服务的一种重要载体，高文、黄铁军指出“Digital Library”的英文本意更强调的是“资料库”，而不是“图书馆”，数字图书馆是一个宽带多媒体网络和海量信息管理系统，其所面对的存储对象和技术领域远远超出了目前传统图书馆的范围[2]。在数字图书馆体系中，由科学家在科学研究或者实验中产生科学数据，由于项目或者共享的需要汇交到专业领域的数据中心，数据中心负责对该数据进行长期保存和服务，并从专业的元数据信息中抽取出符合数字图书馆目录中需要的信息，图书馆将通过最简的元数据集合快速构建联合目录，成为科技信息资源快速的发布门户，便于用户进行检索。

### 1.3 数字图书馆中科学数据目录体系将有利于实现科学数据共享

科学数据目录是实现科学数据资源对外提供服务的重要手段，科学数据目录有助于实现科学数据资源的共享，进一步提升科学数据的应用价值，将科学数据资源纳入目前已经相对成熟的数字图书馆系统中，将有效地整合离散的科学数据资源，构建网络化的科学数据管理和共享服务体系，快速发现和定位科学数据资源，展示数据与文献的关联信息，从而扩展数字图书馆资源服务的领域和深度。

## 2 国内外研究现状概述

面向内容和资源是数字图书馆的核心，在传统文献资源的基础上，越来越多的图书馆机构开始重视非文本（non-texture）资源的建设，图书馆和科技信息服务机构正在面向服务对象，拓展传统的文献服务领域，广泛集成各种数据资源，同时收集科学数据、文献和网络信息等资源，并探索建立不同类型信息资源间的关联。

### 2.1 德国科技图书馆依托DOI技术整合科学数据资源，提供联合目录检索服务

作为科技信息的重要保藏和服务机构，德国科技图书馆成立了非文本资源能力中心（non-texture competence centre），专门搜集整理非文本信息资源，包括视频、音频等科学数据资源，并提供了科学数据资源的目录服务，各个数据中心积极汇交各自的数据资源目录，截止到2007年10月份，在TIB已经通过DOI这种方式注册了475,276个数据集、12,546个科学电影剪辑、6302个医学案例、342个技术报告和112个学习对象等，并部分实现了科学数据与科技文献的引用关联[3]。2009年由该机构联合全球十几家科技信息机构成立了非营利性组织Datacite，为科学数据集提供唯一标识符和公共登记系统，支持数据集的规范引用和复用，并纳入CrossRef系统与文献的链接；多家出版商也发起了Dryad项目，对科学期刊文章中引用的科学数据集进行登记、描述、保存和公共获取服务，有利于构建全面的科学数据发现、关联、利用和复用的基础环境。

### 2.2 国际科技信息机构开始建设全国性的科学数据中心

一些国家的图书情报研究机构作为独立的第三方已经开始着手开展科学数据目录体系的建设。韩国情报所在2012年新成立了韩国国家科学数据中心，制定了长达10年的科学数据中心发展规划，将整合和集成全国重要的科学数据资源，为科学数据的计算、存储提供应用环境，来应对大数据时代面临的数据爆炸却难以有效管理的问题。另外，加拿大国家情报所成立了科研数据管理中心（research data repositories），将开发面向普通用户需求的科学数据目录服务。

### 2.3 国内图书馆跟科学数据中心合作提供简单的科学数据目录服务

中科院国家科学图书馆提供的跨界检索服务系统是由国家科学图书馆自主开发的、面向非文献型信息资源的集成检索服务平台，该平台上能实现对国内外重要科学数据中心的数据链接服务。“中国（西部）环境与生态科学知识积累平台”是由国家科学图书馆兰州分馆与

中国西部环境与生态科学数据中心合作共同建立的一个领域知识平台。该平台的内容主要包括：重大研究计划产出的科学论文、专著、报告（包括演示文件）、野外考察资料、照片和视频等；有关的科学文献；有关的科学数据库。提供文献资源与数据资源的统一检索服务。在该平台中对每条科学数据的描述信息中增加了相关文献的内容，这些文献主要包括两种类型：一种是由该数据直接产生的成果，另一种是对该数据进行补充说明的文献[4]。

## 3 科学数据目录体系建设方案探讨

科学数据目录的概念类似于图书馆中使用的分类目录（catalog）的概念，是以核心元数据为主要描述方式，按照科学数据资源分类体系或者其他方式对科学数据目录资源的有序排列，同时通过科学数据与科技文献的有机联系，通过科学数据目录能够准确地了解和掌握科学数据资源的基本概况，发现和定位所需要的科学数据资源，并能快速地链接到关联的科技文献信息。目录体系建设和使用过程中存在三类角色：信息提供者、信息使用者和目录信息管理者。信息提供者负责信息的编目、注册，保证编目信息的正确性和实时性。信息使用者通过网站和应用系统查询公共资源目录和交换服务目录，发现所需信息。目录信息管理者负责资源目录的建立和管理，并保证目录信息的安全和维护。

数字图书馆中科学数据目录要实现3C（Collection，Classification，Cataloging）的功能，也就是对数据进行收集、分类、编目，这是一个目录体系所要具备的最基本的功能。对于分类和编目，第一是数据目录系统，在数据收集整合的基础上，制定稳定全面的数据分类系统，对数据集进行分类，方便用户迅速定位到数据集；第二是数据检索系统，对数据进行检索和查询，用户能获得相关的信息。第三是特色数据集的建设，表现为数据以主体库的方式进行组织，方便专业研究人员使用。三个层面的数据获取工具相辅相成，为用户提供服务。

科学数据目录体系建设主要由标准规范、技术支撑平台、管理机制构成。

### 3.1 标准规范

科学数据目录体系的标准规范是建立科学数据目录体系的核心，要在科学数据资源的分类方法、元数据、编码规则、标识语言、数据格式、交换协议、资源组织、管理结构等方面制定一系列的标准规范[5]。

科学数据共享系统在构建的时候经常通过扩展通用元数据的方法来创造自己特有的元数据标准，中国的元数据标准扩展多基于ISO、OGC或FGDC的元数据标准，中国可持续发展信息元数据标准草案、科学数据库元数据标准（SDBCM）也是参考了已有元数据的标准构建的，此外，一些领域性较强的科学数据如自然科技资源共享平台构建了符合资源本身特性的元数据描述框架。

数字图书馆的科学数据目录体系的建立是为了帮助用户在数字图书馆的统一平台上获取更多的信息、数据为目的，因此该平台只保存科学数据的目录信息，而科学数据实体保存在各科学数据机构的数据库中，且该目录信息是在科学数据元数据的基础上，参考DC等数字图书馆通用元数据标准，建立科学数据元数据与DC元数据的映射关系，形成数字图书馆中科学数据的核心、最简单的元数据集合，最大程度上满足了科学资源描述和定位的要求；该资源通过对已有的科学数据资源网站进行自动化的抽取获得，通过标准间的映射关系，能方便地与数字图书馆中其他类型资源进行关联和集成。

数字图书馆中科学数据目录核心元数据包括7个必选的元数据实体和元数据元素及1个可选的元数据实体，7个必选元数据实体分别是资源名称、资源摘要、资源提供方、资源分类、资源标识符、资源发布日期、资源分类，并且能与与都柏林核心（DC）标准以及文献的元数据进行初步的对应，便于建立科学数据资源与数字图书馆中其他类型资源之间的映射关系，从而实现联合目录，如表1所示。

### 3.2 技术架构

数字图书馆中科学数据目录将基于数字图书馆框架进行构建，并能方便与现有图书馆目录体系进行综合集成和互操作，从技术角度划分为四个子系统：编目子系统、目录报送子系统、目录管理子系统和目录服务子系统[6]，该体系中涉及的关键技术主要包括元数据技术、数据分类技术、唯一标识符技术等，其技术架构如图1所示。

（1）科学数据编目子系统

基于科学数据核心元数据标准开发的元数据生成工具，从不同科学数据机构的元数据中抽取出符合科学数据目录的元数据信息，生成各自领域的科学数据目录。其主要功能如下：

表1 科学数据目录体系核心元数据规范

| 科学数据元数据字段名称 | 必选/可选 | 与DC的对应 | 与文献的对应 |
|---|---|---|---|
| 数据唯一标识符Identifier | （M） | DC：identifier | 唯一标识符Identifier |
| 数据集名称Title | （M） | DC：title | 文献题名Title |
| 数据发布日期Publish_date | （M） | DC：date | 文献发表时间Publish_date |
| 数据摘要Description | （M） | DC：Description | 文献摘要Abstract |
| 数据贡献者Contributor | （M） | DC：Publisher | 文献作者Author |
| 关键字说明Keywords | （M） | DC：Subject | 文献主题词Keyword |
| 数据分类Category | （M） | DC：Type | 文献分类 |
| 在线资源链接地址URL | （O） | None | 文献链接地址URL |

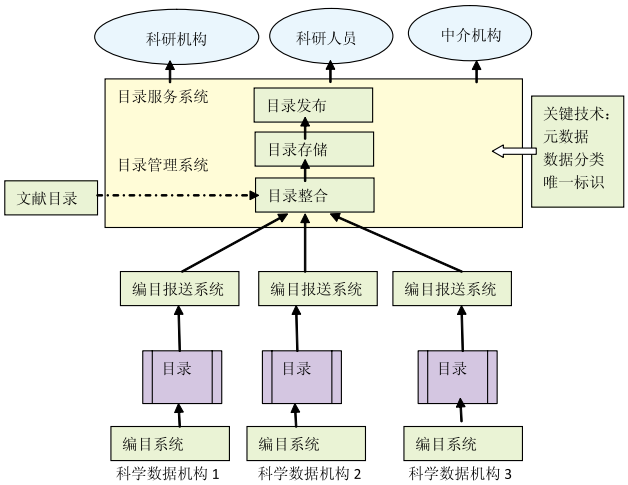


图1 科学数据目录体系技术架构图

元数据生成：针对已有目录，采用不同元数据标准转换方式生成符合科学数据目录的元数据；针对网页、数据库、卫星数据等，可开发专门的元数据抽取工具对元数据进行加工，或者通过在线收割的方式快速获取科学数据目录；对于数字化程度不高的科学数据资源，也可利用手工方式进行元数据录入。

科学数据编码：使用资源编码的前段码并赋予后段码，生成资源ID。目前可采用数字对象唯一标识技术（DOI）对科学数据资源进行编码，用一组数字或字符来实现对科学数据资源的唯一标识，其中每个科学数据机构将获取IDF授权的唯一的DOI前缀，后缀部分将在代理和科学数据机构之间协商产生，采用分段有意义码对科学数据资源进行唯一标识。

科学数据资源分类：目前科学数据常见的分类主要包括使用线分类法、面分类法和混合分类法，以及基于主题词表的分类方法、固定的标准分类方法等[7,8]，科学数据共享工程的分类体系，将科学数据分成了资源环境、基础科学、农业科学、区域综合、人口健康和工程技术四大类，在大类以下进行二级分类，以保持科学数据资源分类的科学性和系统性[9]。在具体的分类过程中，科学数据目录体系中的各个分中心及参与方各自建立自己学科和领域内的分类体系，根据其地域和学科特色对所属的数据进行独立的分类。当数据资源汇交到总中心以后，管理员对其进行分析后和归类，将其映射到总中心的分类体系中。

（2）目录报送子系统

数字图书馆科学数据目录体系利用图书馆目录中心和科学数据分中心之间的专网来实现元数据报送。数据报送系统的功能主要将各科学数据机构的目录元数据，按照科学数据目录体系元数据规范的要求，报送到图书馆目录服务中心。

（3）目录管理子系统

目录管理系统包括数据互操作、元数据管理和系统管理模块，将来自不同来源的科学数据目录系统以及科技文献目录资源进行整合，并按照数字图书馆目录体系标准进行集中存储，实现对目录数据服务的集中管理。

数据互操作模块：按照元数据标准来对外提供目录数据发布和目录数据注册服务。

元数据管理模块：用于管理科学数据元数据的注册、更新、删除，并担负元数据有效性检查的工作。

系统管理模块：实现数据的备份、迁移、认证、用户管理等功能，是提高系统运行可靠性和提高系统可维护性的关键[10]。

（4）目录服务子系统

目录发布子系统将科学数据元数据按照多种分类方式发布到数字图书馆总中心网站，以多种查询方式向用户提供目录服务，供使用者（包括科研机构、科研人员和中介服务机构）进行浏览、查询。

## 3.3 管理机制

科学数据目录体系的管理机制是保证目录体系能够持续、有效运行的一系列管理要求、操作规范和评估机制[11]，主要包括对信息资源、技术平台、业务服务的管理规范，其中对科学数据目录信息维护的管理机制主要包括保证信息采集的持续性、正确性、一致性等管理规范；建立科学数据资源采集、组织、分类、保存、交换、发布与服务管理制度；建立科学数据资源分级联合编目、申报与等级制度；建立科学数据资源唯一标识申请、分配制度。

## 4 结语

数字图书馆的目标就是要建立统一的资源目录系统，并在系统中实现多种资源的集成和应用，科学数据作为一种重要的信息资源，其重要性在科技融合的大背景下显得尤为重要。但由于科学数据类型多样、结构复杂，其目录体系的建设将是一个长期艰巨的过程，并需要在长期与文献资源的整合、引用的过程中，最终实现科技信息资源的全面整合与共享。

**参考文献**

[1] 李晓波.科学数据共享关键技术[M].北京:地质出版社,2007.
[2] 高文,刘峰,黄铁军.数字图书馆:原理与技术实现[M].北京:清华大学出版社,2009.
[3] BRASE J.德国国家科技图书馆科学内容DOI注册中心研究进展[J].中国科技资源导刊.2008,40(1):37-39.
[4] 祝忠明,马建霞,常宁,等.SEEKSpace——基于Dspace的环境与生态科学知识积累平台[J].图书情报工作.2007,51(4):71-74.
[5] 吴晓敏,刘小白.政府信息资源目录体系建设初探[EB/OL]. [2012-07-13]. http://www.ciotimes.com/information/topic/topic200806101514/index.htm.
[6] 王卫文,谢先江.电子政务信息资源目录体系构建的研究[J].现代情报,2006(7):219-222.
[7] 耿庆斋,张行南.基于多维组合的水利科学数据分类体系及其编码结构[J].河海大学学报(自然科学版),2009,37(3):346-350.
[8] 廖顺宝,蒋林.地球系统科学数据分类体系研究[J].地理科学进展,2005,24(6):93-98.
[9] 中华人民共和国科学技术部基础研究司.数据分类与编码的基本原则与方法[M].中华人民共和国科学技术部,2005.
[10] 谢先江.区域性电子政务信息资源目录体系实现研究[J].电子政务,2007(12):37-41.
[11] 张乃丁,李刚.城市建设信息资源目录体系构建研究[EB/OL]. [2012-07-13]. http://www.ciotimes.com/information/topic/topic200806101514/index.htm.

**作者简介**

涂勇（1981-），男，湖北鄂州人，博士，中国科学技术信息研究所助理研究员，研究方向为科技资源管理、科学数据共享、数字对象唯一标识等。
E-mail: tuyong@istic.ac.cn

## Discussion on Program for Scientific Data Catalogue System in Digital Library

Tu Yong, Peng Jie / Center for resource sharing promotion, Institute of Scientific and Technical Information of China, Beijing, 100038
Guo Xiaofeng / Beijing Wanfang Data Co., Ltd., Beijing, 100038

Abstract: The necessity of the construction of scientific data catalog system in digital libraries is analyzed in the paper, and the concept of scientific data catalog system is forwarded on the basis of comparative study of scientific data directory system development in domestic and foreign scientific and technological information agencies. Finally, three aspects, which is standards, technical structure and management mechanisms, on scientific data catalog system framework in digital library are described.

Keywords: Digital library, Scientific data, Catalogue system

（收稿日期：2012-07-13）