doi:10.3772/j.issn.2095-915x.2020.01.003

国际主要科学数据集检索平台对比研究

杨波 赵扬 焦红

南京农业大学 南京 210094

开放科学（资源服务）标识码（OSID）

摘要：在开放科学运动蓬勃发展的背景下，开放共享理念已成为科学发展和知识传播的基础，科学数据的有效组织和共享是其重要的研究课题之一。为了帮助科学家缩短科研活动中的数据获取周期、降低数据获取的难度并提升数据复用的效率，本文将从检索入口、收录范围、核心功能等多个角度，深入比较国际五大科学数据集检索平台的优劣。经过对比分析发现，无论是综合性平台还是聚焦于特定领域的专业性平台，在系统功能和数据集收录方面都展现出了不同的特点。虽然尚处于开发和应用的初级阶段，但是这些平台依托于各自的技术和资源优势，均以各种不同的方式，初步实现了科学数据集的组织和共享。

关键词：科学数据管理；科学数据集；数据共享

中图分类号：G250；G350

Comparison Study of International Major Scientific Dataset Retrieval Platforms

YANG Bo ZHAO Yang JIAO Hong

Nanjing Agricultural University, Nanjing 210094, China

Abstract: In the context of vigorous advance of open science movement, the idea of opening & sharing has become the basis of development of science and knowledge dissemination, among which, effectively organizing and sharing scientific data is one of the key topics. In order to help scientists to shorten the cycle of data acquisition, reduce the difficulty of data acquisition and improve the efficiency of data reuse in scientific activities, this paper will make an in-depth comparison of the advantages and disadvantages of five international retrieval platforms for scientific data sets from the perspectives of access entry, inclusion scope and core function. It is found that both the comprehensive platforms and some professional ones focusing on a specific field present different characteristic in terms of system functions and resources of datasets. Although most of the platforms are in the

基金项目：国家社科基金一般项目“科学数据集的自组织模式和质量评价研究”（18BTQ007）。

作者简介：杨波（1981-），教授，博士生导师，研究方向：Web 结构挖掘，科学数据管理，基于大规模数据集的科技评价，E-mail: boyang@njau.edu.cn；赵扬（1996-），硕士研究生，研究方向：信息计量；焦红（1991-），博士研究生，研究方向：信息计量。

initial stage of development and application, relying on their advantages in technology and resources, they have basically fulfilled the task of the organization and sharing of scientific datasets in various ways.

Keywords: Scientific data management; scientific dataset; data sharing

引言

随着科学数据的不断增多，对数据进行有效组织和共享已成为科学数据管理的重要研究课题。近年来，多个著名的信息服务商开发了包含各种形式、覆盖不同学科的科学数据集检索平台，为科学数据的组织、检索、质量评价和共享提供了极大的便利。由于科学数据的共享机制尚不成熟，大多数科学数据集检索平台的开发和服务模式仍处于探索阶段。从单一主题的角度、以集合的方式对科学数据进行组织，是当前学术界在数据存储、共享和评价以及成果展示等科研活动中最为主要的形式之一。为了提高研究人员整合数据资源的效率，本文通过分析当前国际主流科学数据集检索平台的，对照检索服务等基本指标，综合评价检索平台，为进行数据集检索工作的研究人员提供参考。

1 科学数据集检索平台简介

搜索引擎是一种为用户提供 Web 信息查询服务的软件系统，它以一定的策略在 Web 上搜集和发现信息并对信息进行处理和组织，最终返回与用户查询信息相关的结果列表 [1]。学术搜索引擎则是以各类学术资源为索引对象的一种搜索引擎，常见的学术搜索引擎有 Google Scholar、Scirus、百度学术等。科学数据集检索平台与学术搜索引擎都能帮助科研人员获取科研活动中所需的相关资源，但二者针对的资源对象不尽相同。传统学术搜索引擎整合了期刊论文、专利、图书、报纸等不同类型的文献，而科学数据集检索平台则索引了不同类型的数据集。根据学术搜索引擎的相关定义，本文将科学数据集检索平台定义为：通过一定的策略索引不同类型的数据集，从而为科研人员提供数据集查找、获取、分析等服务的平台。为了展示各平台的不同特点，本文选取了部分国际主流的科学数据集检索平台进行对比研究，以下对其相关信息进行简要介绍。

（1）Data Citation Index（简称 DCI）

DCI 是科睿唯安公司（原汤森路透）于 2012 年推出的科学数据集引用统计和共享平台，提供了包含多学科、链接多元知识库的高质量研究数据 [2]。由于依托于著名的文献数据库 Web of Science，DCI 一经发布就引起了学术界的广泛关注和使用。

（2）DataCite Search

DataCite[3] 是一个全球性的非营利性组织，该组织提供科研成果（包括科学数据）的 DOI 注册服务 [4]，为科学数据集的共享提供了极大的便利条件 [5]。DataCite 要求用户在注册数据集时，必须使用其提供的元数据架构（Metadata Schema）[6] 对数据集进行描述，并将元数据上传至元数据存储库（Metadata Store）[7]。正是因为 DataCite 拥有大量具有唯一标识符的数据集资源，DataCite 于 2015 年 8 月 17 日推出了

doi:10.3772/j.issn.2095-915x.2020.01.003

一项数据搜索工具 DataCite Search[8]。DataCite Search 通过专门为科学数据集的元数据建立索引的方式，为用户提供跨库查找数据集资源的一站式数据检索服务。

（3）DataMed

DataMed 是由美国国家卫生研究院（NIH）资助的 BD2K 项目（Big Data to Knowledge）开发的一个生物医学搜索引擎，已于 2016 年 6 月 30 日发布了 1.0 版本 [9]，其目标是从不同的数据存储库或数据聚合器中发现生物医学数据并提供相关的访问信息 [10]。

（4）DataSearch（测试版）

由全球著名的期刊出版商 Elsevier 开发的数据搜索引擎 DataSearch 于 2016 年 6 月发布了第一个测试版。DataSearch 允许用户跨领域和跨数据存储库查找不同类型的数据资源 [11]。除了索引开放获取库（Open Access）中的核心数据集，DataSearch 还索引来自 Elsevier 或开放获取库的图像、表格和补充数据 [12]。

（5）Dataset Search

Google 于 2018 年 9 月 5 日推出了 Dataset Search 的测试版，经过一年多的测试与用户反馈后，2020 年 1 月 24 日推出 Dataset Search 正式版本 [13]。Dataset Search 允许任何人搜索发布在公开媒体，如网站、数字图书馆和个人主页上的数据 [14]，数据提供者也可以采用结构化数据标准格式来描述其数据，以便被 Dataset Search 收录 [15]。

五大科学数据集检索平台的基本信息如表 1 所示。

表 1 检索平台基本概况

名称	隶属机构	平台地址	上线时间	平台性质
DCI	科睿唯安	通过机构订阅后访问	2012年	付费
DataCite Search	DataCite	https://search.datacite.org/	2015年	免费
DataMed	美国国家卫生研究院	https://www.datamed.org/	2016年	免费
DataSearch（测试版）	Elsevier	https://datasearch.elsevier.com/	2016年	免费
Dataset Search	Google	https://datasetsearch.research.google.com	2017年	免费

2 科学数据集检索平台对比

随着数据驱动的科学研究模式日益成熟，科学数据作为一种重要的学术资源和成果形式，得到了学术界的广泛关注和认可。科学数据与科学论文密切关联，以数据提及（Data mentions）、数据引用和数据发表为主要表现形式的数据共享需求，成为各大文献数据库共同认可的新型文献服务增长点，使得新兴的互联网文献数据库商也涉足其中，纷纷开发了自己的数据集检索平台。然而，这些平台依托于不同的文献库资源（学术论文或网络文献），所采用的技术路线、服务模式和服务对象并不完全相同。因此，本节将从访问入口、数据收录概况及核心功能等方面对比和分析上节所提及的国际五大科学数据集检索平台。

2.1 访问入口功能比较

相比于科学文献数据库中文献资源的规范性，不同科学数据集的来源、格式、大小、结

构和隶属机构等属性，以及科学数据集检索平台的设计理念、服务对象、收录范围存在巨大差异。因此，大多数平台的访问入口设计均有所不同。当前主流的科学数据集检索平台访问入口的功能主要集中在关键词检索、分类浏览、概况展示和系统帮助等方面。尤其是后两者，对于用户及时了解该平台数据量、收录范围、数据类型等信息，帮助用户快速掌握系统使用方法，从而找到最有价值的科学数据集来说，是至关重要的。

五大平台访问入口功能的比较如下表 2 所示。

表 **2** 访问入口功能

功能列表 / 平台名称	关键词检索	分类浏览	索引数据概况	用户注册/登录	使用帮助	界面语言
DCI	√	×	×	√	√	多语言
DataCite Search	√	√	×	√	√	英文
DataMed	√	×	√	√	√	英文
DataSearch（测试版）	√	×	×	×	√	英文
Dataset Search	√	×	×	√	√	多语言

关键词检索与使用帮助功能是检索平台应具备的基础功能，前者能提高平台的易用性，后者则有助于用户找到解决问题的途径。由表 2 可知，五大平台在访问入口处都提供了这两种功能。

分类浏览功能既能帮助用户了解平台数据索引的组织概况，也有助于用户的分类查找。在分类浏览方面，只有 DataCite Search 拥有该功能。DataCite Search 支持对数据集、收录的存储库以及 DataCite 成员等信息进行分类浏览，并且会显示某一类别下的具体数据量以及数据摘要。在数据集分类浏览页面，平台提供了注册年份、隶属机构、资源类型等限定属性，允许用户在关键词检索的基础上，利用这些属性进行二次筛选。

索引数据概况的展示可以帮助用户了解该平台索引的数据量、数据库数量以及数据类型等信息。一方面，用户可以初步判断该平台是否能找到所需数据；另一方面，用户也可以知道相应的检索字段，以便后续查找。在索引数据概况方面，只有 DataMed 在首页展示了索引的科学数据集来源（存储库）、数据类型、数据集的数量和前沿研究等统计信息，用户可通过点击统计信息的相关链接进一步了解具体的科学数据集检索平台情况。除此之外，DataMed 还对数据量最大的前 8 个存储库的收录详情，以及最受欢迎的七个数据集的信息进行了突出展示。

注册与登录是平台提供用户个性化服务的前提，只有提供注册服务，才能在登录后为用户提供保留检索、浏览历史等服务。由表 2 可知，除了 Elsevier 的 DataSearch 之外，其余四大平台都具备该功能。其中 DCI 支持用户以个人身份注册登录，也可通过机构账号登录使用。在 Dataset Search，可以使用谷歌账号登陆该平台。DataMed 支持邮箱注册登陆和谷歌账户登陆两种方式。Datacite Search 不但支持邮箱注册、机构登录和谷歌账户登陆，用户还可以使

doi:10.3772/j.issn.2095-915x.2020.01.003

用 ORCID 账号登录平台。当用户通过个人邮箱注册登录时，DataCite Search 提供关联 ORCID 与当前帐户的功能。

平台的语言切换功能可以为不同国家的用户提供语言便利。由表 2 可知，只有 DCI 和 Dataset Search 支持多种语言切换，其他平台均为英文界面。其中 DCI 支持中文、英文、日文等 7 种语言，Dataset Search 则支持英文、中文、日文、韩文等 18 种语言。

2.2 数据集收录情况比较

在进行科学数据集检索时，检索平台索引科学数据集的学科范围、科学数据集数量、数据类型等要素是用户选择的重要依据，也是评价各种平台的数据可用性和研究适用性的重要参考。

根据表 3 的统计数据可知，在覆盖学科范围方面，其他四大平台的学科覆盖范围都较为全面，而 DataMed 索引的科学数据集只涉及生物医学领域。这与不同平台的隶属机构及开发目的相关。在索引数据集数量方面，可以看出 Dataset Search 索引的数据集量最多，达到 2500 万。与其他平台不同的是，Google 推出的 Dataset Search 不是通过索引数据集存储库来提供科学数据集查找服务，而是对网络上所有使用开放标准（如 schema.org 和 DCAT）描述的科学数据集进行抓取，并收集其关联的元数据 [16]，因此其索引数据集数量较多。DataCite Search 索引数据集数量位居第二，这与其提供的 DOI 注册服务是分不开的。DCI 的数据集和数据库的数量都较多，与其提供服务时间较早，以及拥有强大的论文库做支持有一定的关系。而 DataMed 虽然只索引了生物医学领域的科学数据集，但是其数据量仍有 233 万。DataSearch 由于还处于初步的测试阶段，其索引的数据量还未可知。

表 3 科学数据集收录概况

指标 \ 平台名称	DCI	DataCite Search	DataMed	DataSearch（测试版）	Dataset Search
学科范围	综合性	综合性	生物医学	综合性	综合性
索引数据集数量	1021万	1841万	233万	未知	2500万
索引数据库数量	414个	1877个	75个	18个	未知

了解不同检索平台的数据集分类方式，可以帮助用户更好地进行分类查找，同时可以为用户查找特定格式的数据集时提供依据。在数据集分类方面，不同检索平台对科学数据集的分类方式各不相同。DCI 先从文献类型的角度将数据集分为 Repository（知识库）、Data study（数据研究）、Dataset（数据集）和 Software（软件），在不同的文献类型下又对科学数据集类型进一步划分。除 DCI 外的其他四个平台，首先从科学数据集的分类数量上来看：Dataset Search 官方显示其索引的具体科学数据集类型有 7 种，DataSearch 索引的科学数据集类型有 12 种，DataCite Search 与 DataMed 索引的科学数据集类型都为 15 种。其次从科学数据集的分类类型上来看：由于主要关注生物医学领域的科学数据集，因此 DataMed 从生物医学的角度

对科学数据集进行分类，其他平台则是以数据资源的文献类型分类为主。总之，除了上述提到的较为特殊的 DCI 外，其余四个平台的科学数据集分类体系比较相似，详情如表 4 所示。

表 4 科学数据集分类体系

平台名称	DataMed	DataCite Search	Datasearch（测试版）	Dataset Search
数据集类型	蛋白质	数据集	图片	表格或CSV文件
	表型	图片	表格数据集	表格集
	基因表达	文本	文档	专有格式文件
	核苷酸序列	音频	视频	文件集
	形态	软件代码	文件集	结构化数据集
	临床实验	模型	文本	图像捕获文件
	蛋白质组学数据集	互动性资源	软件/代码	机器学习有关文件
	生理信号	物理对象	幻灯片	其它
	表观遗传数据集	服务	地理空间数据集	
	论文数据集	活动	音频	
	组学数据集	工作流	测序数据集	
	调查数据集	视频	其它	
	细胞信号传导	视听资源		
	图像数据集	集合		
	其它	其它		

2.3 核心功能比较

作为提供科学数据集检索服务的平台，科学数据集查找是评价平台易用性的核心环节之一。科学数据集查找可划分为三个阶段：数据集检索、结果输出和个性化服务。因此，本节将从这三个阶段出发对五大科学数据集检索平台的功能进行比较。

2.3.1 检索功能比较

检索阶段主要从检索方式、检索字段、检索记录保留以及错误提示这四个方面进行比较，主要的检索功能如表 5 所示。

表 5 检索功能概况

平台名称 / 检索阶段	DCI	DataCite Search	DataMed	DataSearch（测试版）	Dataset Search
检索方式	基本检索、被引参考文献检索和高级检索	全文检索、分类检索和高级检索	全文检索、分类检索和高级检索	全文检索、高级检索	全文检索、高级检索
检索字段个数	15个	19个	21个	6个	未知
检索记录保留	支持	不支持	支持	不支持	不支持
错误提示	支持	不支持	支持	不支持	不支持

从检索方式来看，五大平台都支持高级检索。其中 DCI 支持施引文献检索，从而可以通过施引文献了解某一科学数据集的被引情况。除 DCI 外，其他四个平台都支持全文检索。只有 DataCite Search 和 DataMed 支持分类检索，但又各有不同：DataCite Search 的分类检索与平台入口的分类浏览功能保持一致，而 DataMed 的分类查找分为按数据查找和按数据所在存储

doi:10.3772/j.issn.2095-915x.2020.01.003

库查找。

从检索字段来看，不同的检索平台提供的检索字段各不相同。除了 Dataset Search 未明确指出可供检索的字段外，其余四个平台均提供标题、作者和出版年份的检索。其中，DataMed 提供的检索字段最多（21 个），DataSearch 提供的检索字段最少（6 个）。

检索记录的保留可以帮助用户记录检索历史，便于数据回溯和记录合并。DCI 和 DataMed 集成了该功能，前提是用户必须登录系统。

用户在检索时往往会出现各种错误，比如拼写错误、检索用词不够准确以及检索词组配的语法错误等。因此，作为一个提供数据集查找服务的平台，在用户检索过程中及时提示错误或者进行同义词推荐是十分重要的。五大平台中只有 DCI 和 DataMed 具备在检索过程中及时提示错误，或者推荐同义词的功能。当用户在检索框中输入检索词或检索式后，DataMed 会在网页右侧进行同义词推荐，并且会根据当前的输入进行检索词的自动扩展。

2.3.2 结果输出比较

用户检索后，如何在结果页面快速查找到所需的相关数据集，这就依赖于平台所提供的结果输出服务。在结果输出阶段，主要从结果统计、翻页、排序、分类筛选和相关推荐这五方面对五大平台进行比较。具体对比结果如表 6 所示。

表 6 结果处理功能

平台名称 / 结果输出	DCI	DataCite Search	DataMed	DataSearch（测试版）	Dataset Search
结果统计	总记录数、分类记录数和每页记录数	总记录数和分类记录数	总记录数、分类记录数、每页记录数和当前页数	总记录数和分类记录数	总记录数
每页结果数	10/25/50	25	5/10/20/50/100	10	不支持翻页
结果筛选	支持	支持	支持	支持	支持
结果排序	支持	不支持	不支持	不支持	不支持
相关推荐	不支持	支持	不支持	不支持	不支持

访问各大平台后发现，五大平台在结果返回页面都以数据摘要的形式展示数据集列表。其中比较有特色的是 DataSearch 和 Dataset Search。在 DataSearch 的结果页面，用户可以通过点击某一具体数据集摘要来查看数据集的详细信息（如发布时间、作者等）和数据集包含的数据文件，如图 2 所示。Dataset Search 则是采取分栏显示的方式展示结果细节（如图 3 所示）。

由表 6 可知，五大平台都对返回结果进行了统计。Dataset Search 没有显示返回结果数量的确切数字，如图 3 左侧显示的“找到 100 多个数据集”。其他四大平台在结果输出页面都对检索结果总数进行了准确显示，以 DataCite Search 为例（图 4）。DataCite Search 和 DataSearch 每页显示结果数分别为 25 条和 10 条，DCI 和 DataMed 则可设置每页显示的结果数。

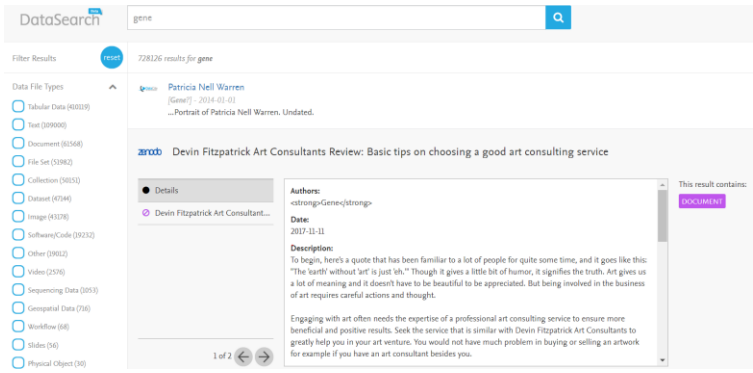

图 1 **DataSearch** 结果查看图

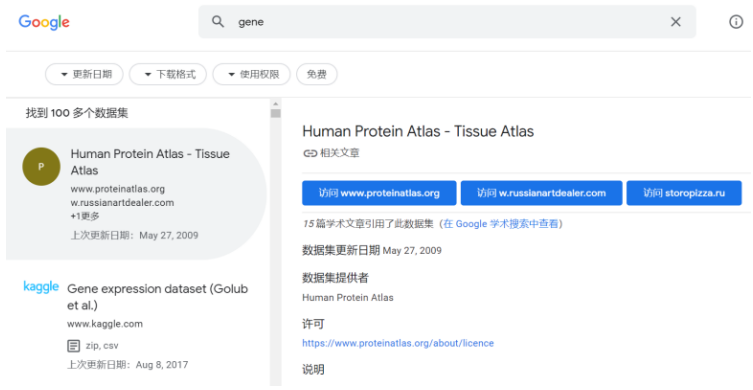

图 2 **Dataset Search** 结果输出页面图

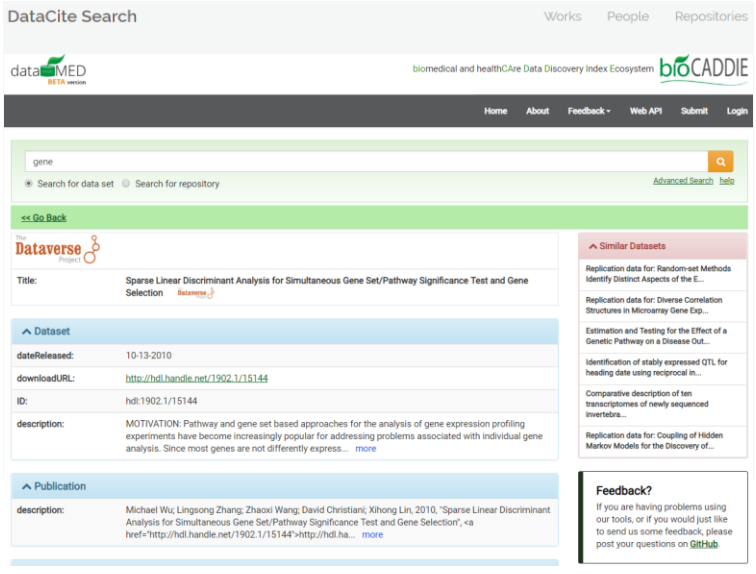

图 4 **DataMed** 搜索结果页面

在结果排序方面，只有 DCI 支持结果排序功能。DCI 从出版年、被引频次、使用次数和相关性等 8 个方面对结果进行排序。用户通过选择输出结果的排序方式，不但可以根据出版日期降序了解最新的科学数据集信息，而且可以通过相关性排序快速定位所需

doi:10.3772/j.issn.2095-915x.2020.01.003

科学数据集。

在结果筛选方面，除了 Dataset Search，其余四大平台都按不同的分类字段对检索结果进行了统计。具体分类筛选字段详情如表 7 所示。

表 7 结果分类筛选对比

平台名称	DCI	DataMed	DataCite Search	Datasearch（测试版）	Dataset Search
结果分类筛选字段	出版年 Web of Science类别 文献类型 科学数据集类型 来源机构 作者 编者 团体作者 国家/地区 来源存储库 语种 学科类别	数据集类型 来源存储库 可获得性 数据集授权	注册年份 数据集类型 来源机构	数据集类型 来源存储库类型 来源存储库 发布日期	更新日期 下载格式 使用权限 是否免费

由表 7 可以看出，DCI 的分类筛选选项最多，灵活性相对更好。并且由于 DCI 引入了 Web of Science 的分类体系，因此在数据和文献的一体化协同分析方面独具优势。相比之下，其余四个平台分类筛选字段较少。其中 DataMed、DataCite Search 和 DataSearch 主要是从数据集的类型、来源以及发表时间这三方面对返回结果进行分类。Dataset Search 除了数据集的更新日期字段，其他字段都是从用户角度出发对数据集进行分类。

相关推荐功能可以根据用户搜索某一数据集的行为，自动推荐该数据集的不同版本或与之相关的其他数据集。不同版本数据集的推荐可以帮助用户了解该数据集的版本演变，而其他相关数据集的推荐则是通过相关反馈帮助用户找到所需的数据集。五大平台中，只有 DataMed 提供该功能。详情页的右侧展示了 DataMed 根据搜索行为推荐的相似数据集，如图 4 所示用户可以点击链接查看相似数据集的详细信息。

2.3.3 个性化服务比较

在个性化服务方面，除了用户的注册、登录功能之外，相关数据集的收藏、输出结果分析、引文追踪等也有助于研究人员查找数据集。五大平台中，DCI 和 DataMed 表现优异。DataMed 可以保留已注册用户的搜索查询，并记录用户近期的使用行为。对于感兴趣的科学数据集，登录用户既可以收藏，也可以通过电子邮件分享。用户在注册和登录 DCI 后，可以进行个性化设置，如时间跨度，使用的检索语言等。除此之外，DCI 也支持对检索查询的保存，以便用户在检索历史中查看先前的查询，进行检索式的组配或检索结果的合并。DCI 支持将查询结果导入到 EndNote 账户或电子邮件中，方便用户在后续研究中使用。用户也可以利用 DCI 的检索结果分析功能，DCI 支持从研究方向、出版年等 16 个维度对结果进行分析，并以可视化图表的方式展示分析结果。

2.3.4 共享方式比较

科学数据集检索平台方便了用户以浏览、

下载和引用等方式利用科学数据集，用户使用科学数据集的行为也进一步促进了科学数据集的共享。因此，灵活的共享方式也是评价科学数据集检索平台的一个重要指标。

表 8 科学数据集共享方式

平台名称 / 共享方式	DCI	DataCite Search	DataMed	DataSearch（测试版）	Dataset Search
元数据下载	√	√	√	×	×
定位原始数据集	√	√	√	√	√
提供引用格式	√	√	×	×	×
施引查看	√	×	×	×	√
数据集分享	×	√	×	×	√

如表 8 所示，在元数据下载方面，DCI、DataCite Search 和 DataMed 支持此功能。其中 DCI 和 DataMed 支持选中记录的导出，DataCite Search 不支持批量下载，只能单独下载当前的某一数据集。DataCite Search 提供三种元数据的下载格式，分别是 DataCite XML、DataCite JSON 和 Schema.org JSON-LD。五大平台均提供原始科学数据集的下载地址，有助于用户获取原始数据。DCI 和 DataCite Search 所提供的标准引用格式，一方面为科学数据集的规范性使用声明提供了便利；另一方面，基于引用机制的共享行为也促进了科学数据集的传播，使得数据集的创建者的贡献更容易得到认可。在施引方面，只有 DCI 和 Dataset Search 提供科学数据集的施引查看。DCI 可以通过点击科学数据集的被引次数链接进而查看施引文献列表，在 Dataset Search 中同样也可以通过 Google Scholar 查看引用科学数据集的文章。在科学数据集分享方面，DataCite Search 和 Dataset Search 支持该功能。用户可以将科学数据集的元数据分享到 Facebook 和 Twitter 等社交媒体，方便用户共享交流。

3 检索实例

为了更直观地展示五大平台的检索效果，本节选取部分中、英文检索词进行检索测试，结果如下表所示。

表 9 检索结果输出

检索词 / 平台名称	rna	基因	protein	线粒体
DCI	469120	0	883062	0
DataCite Search	96776	1253	338123	981
DataMed	207727	2	449032	0
DataSearch（测试版）	260520	0	906891	0
Dataset Search	100	36	100	0

由表 9 可知，不同平台对于同一检索词返回的结果数量各不相同。对于英文检索词，返回结果数较多的平台为 DCI、DataMed 和 DataSearch。Dataset Search 虽然索引数据集的数量最多，但返回的结果数较少。对于中文检索词，各大平台返回的结果都较少，甚至部分平台返回结果数为 0。由此可见，各大平台对于中文数据集的收录方面有待加强。

在查看返回结果的方面，DCI 和 Datacite

doi:10.3772/j.issn.2095-915x.2020.01.003

Search 仅支持查看部分数据集，其中 DCI 最多可查看 10 万数据集，Datacite Search 则只支持查看 1000 条数据集。DataMed 和 DataSearch 均可以通过翻页的方式查看所有返回的数据集。Dataset Search 可通过网页右侧的滚动条查看所有的返回结果，返回结果数若是超过 100 条，页面只会显示 100 条数据集。

4 结语

开放科学运动不仅让科研活动的开展更加便捷、科研资料的共享变得更加普遍，而且推动了科学数据集的共享。长期以来，科学数据资源极低的重复利用率和科研人员日益旺盛的数据需求之间的矛盾一直存在。在数据驱动的科研背景下，该问题显得尤为突出。科学数据集检索平台的出现，有望在很大程度上满足科研人员对科学数据集的使用需求。这些工具不仅能够帮助科研人员更加有针对性地选择科学数据集，而且可用来揭示科学数据集与科学研究之间的内在关系，继而进一步推动开放科学范式下跨学科、跨区域的科研合作。

本文主要从访问入口功能、数据集收录情况和核心功能三个方面，对国际五大科学数据集检索平台提供的检索服务进行对比分析后发现：

（1）在访问入口方面，五大平台均支持关键词检索并提供使用帮助。但不同检索平台在首页功能上存在差异，其中 DataMed 对索引数据概况进行了展示，DataCite Search 支持分类浏览功能。除 DataSearch 外，其余四大平台均支持用户注册与登录。

（2）在数据集收录情况方面，DataMed 收录来自生物医学领域的数据集，其余四大平台收录范围都较为全面。Google 推出的 Dataset Search 索引数据量最多，而以索引数据存储库来提供数据集检索服务的其余四大平台中，DataCite Search 索引的数据存储库数量最多。

（3）综合比较不同数据集检索平台的核心功能后发现，付费综合性的检索平台——DCI 与供免费使用的专业性检索平台——DataMed 能带给用户更好的使用体验。Dataset Search 虽然收录的数据集数量最多，但其提供的数据集查找服务还有待加强。仍处于测试阶段的 DataSearch 不论是在数据集收录方面还是在检索服务方面，都需进一步优化。

根据已有研究梳理了科学数据集的开放共享背景、科学数据集的组织形式和组织技术，分析国际主流的五大科学数据集检索平台，发现当前的科学数据集检索平台存在以下特点：

（1）兼有综合性检索平台和专业性检索平台。国际主要的数据集检索平台的建设正处于全面发展阶段，现如今综合性学科领域和特定学科覆盖的数据集检索平台协同构建，不同类型的平台为用户提供了多领域、多途径的数据集检索途径。

（2）各大检索平台都将其依托的特色资源或者技术应用于构建数据集检索平台。例如，DCI 作为科睿唯安的产品之一，雄厚的文献资源储备以及运营 Web of Science 的成功经验成为其强大后盾；DataCite 将 DOI 注册服务应用于数据集检索平台的构建，帮助平台实现了数据集描述和关联的标准化；Dataset Search 借助于 Google 的海量 Web 采集和索引技术，以及

丰富的 Google Scholar 数据库，成为数据量最大的平台之一。总之，各平台的建设单位都在努力将成熟的技术与经验投入到数据集检索平台的搭建中，这无疑推动了数据集检索平台的快速成长。

（3）数据集检索平台的技术成熟度和服务能力尚处于初级阶段。2020 年初 Dataset Search 才发布正式版本 [13]，虽然相对于测试版，该版本提供更加丰富的搜索结果过滤选项，但功能仍然较为单一。DataSearch 目前也还处于测试阶段。除此之外，大部分平台的科学数据集和科学文献的关联分析功能还比较弱。

综上所述，本研究所介绍的这国际五大数据集检索平台不但为科学数据集的共享和复用提供了重要的途径，其技术方案、分析方法和运行机制也可为其他研究科学数据集挖掘和关联技术、数据成果评价的人员提供参考。在全球科学数据开放共享的背景下，数据共享对于快速高效的解决关键技术研发等问题至关重要。虽然数据壁垒依然存在，开放数据的呼声依然强烈 [17]，但这些平台的建立和发布，吹响了从开放数据到开放科学的号角，更为重要的是，也让更多的科研人员相信：数据开放，未来可期！

参考文献

[1] 李晓明 , 闫宏飞 , 王继民 . 搜索引擎——原理、技术与系统 [M]. 北京 : 科学出版社 , 2004:2-3.

[2] Thomson Reuters. The Data citation index [EB/OL]. [2020-03-20]. http://wokinfo.com/ products_tools/multidisciplinary/dci/.

[3] DataCite [EB/OL]. [2020-03-20]. https://datacite.org.

[4] DataCite. DataCite - Mission [EB/OL]. [2020-03-20]. https://datacite.org/mission.html.

[5] Brase J. Making data citeable: DataCite [M] // Opening Science. Springer, Cham, 2014: 327-329.

[6] DataCite. DataCite Metadata Schema [EB/OL]. [2020-03-20]. https://schema.datacite.org.

[7] DataCite. DataCite MDS API Guide [EB/OL]. [2020-03-20]. https://support.datacite.org/ docs/mds-api-guide.

[8] Fenner M. From Pilot to Service [EB/OL]. [2020-03-20]. https://blog.datacite.org/from-pilot-to- service/.

[9] Machado L O, Diego S. bioCADDIE: Progress and Next Steps [EB/OL]. [2020-03-20]. https://datascience.nih.gov/sites/all/themes/datascience/conferences/2016/Ohno-Machado-Abstract.pdf.

[10] dataMED. OUR MISSION [EB/OL]. [2020-02-04]. https://www.datamed.org/about.php.

[11] Mendeley Blog. Introducing Elsevier DataSearch–Beta Two [EB/OL]. [2020-03-20]. https://blog.mendeley.com/2017/09/18/introducing-elsevier-DataSearch-beta-two/.

[12] DataSearch. Frequently Asked Questions [EB/OL]. [2020-1-10]. https://datasearch.elsevier.com/faq#/.

[13] Noy N. Discovering millions of datasets on the web [EB/OL]. [2020-03-20]. https://www.blog.google/products/search/discovering-millions-datasets-web/.

[14] Noy N. Making it easier to discover datasets [EB/OL]. [2020-03-20]. https://www.blog.google/products/search/making-it-easier-discover-datasets/.

[15] Google Search. Understand how structured data works [EB/OL]. [2020-03-20]. https://developers.google.com/search/docs/guides/intro-structured-data#structured-data-format.

[16] Noy N. Facilitating the discovery of public datasets [EB/OL]. [2020-03-20]. https://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html.

[17] Wu C I, Poo M. Moral imperative for the immediate release of 2019-nCoV sequence data [J]. National Science Review, 2020, 7(4):719-720.