从国际著名引文数据库的比较分析看DISC的建设与发展

杨增秀 张桂玲 杨海超 冯艳君 赵卫华 张欣
（机械工业信息研究院，北京 100037）

摘要：本文选取Scopus、Web of Science、CiteSeerX等国际著名的引文数据库，从文献收录范围、元数据描述、检索功能、统计分析功能4个维度与国际科学引文数据库（Database of International Science Citation，DISC）进行对比分析，总结归纳当下国际先进引文数据库在收录范围、检索途径、分析评价等方面所具有的特点，为进一步完善DISC的建设与服务工作提出相关建议。

关键词：引文数据库；Scopus；Web of Science；CiteSeerX；DISC
中图分类号：G354.4　　DOI：10.3772/j.issn.1673-2286.2020.07.009
引用格式：杨增秀，张桂玲，杨海超，等. 从国际著名引文数据库的比较分析看DISC的建设与发展 [J]. 数字图书馆论坛，2020 (7)：67-72.

国际科学引文数据库（http://disc.nstl.gov.cn）是国家科技图书文献中心（National Science and Technology Library，NSTL）于2006年启动的，以成员单位订购的科技期刊和科学引证关系为基础建设的一个大型外文文献检索服务系统[1]。近年来，DISC在为全国科技界用户提供文献信息保障中发挥了重要作用。经过十余年的发展，为适应引文数据大规模增长的发展趋势，更好地为我国科研人员提供世界科学研究的脉络，为其了解世界科学研究动态提供方便，NSTL高度重视DISC数据库的建设，拟对DISC系统进行全面升级改造。为此，本文选取Web of Science、Scopus、CiteSeerX这3个国际著名引文数据库，在对比分析的基础上，归纳各相关数据库的优势与特色，梳理DISC的差距与不足，为进一步改进和完善DISC的检索与服务功能提出相关建议。

1 4个数据库的基本情况

DISC是NSTL自主研发的一个外文科技文献引文数据库，2007年初投入使用。经过十余年的发展，数据库建设已初具规模。DISC具有一定的文献发现功能，用户可以从来源文献和引文等多种途径检索和浏览文献信息，是目前我国科技界用户可以通过网络免费利用的唯一一个拥有自主知识产权的外文文献引文查询服务系统。系统具有与NSTL文献原文传递和代查代借系统无缝链接的功能，支持用户快速获取文献全文，是用户获取与利用NSTL文献信息服务的一个重要途径[2]。

Scopus（https://www.scopus.com）是由爱思唯尔出版公司研发的全球最大的文摘引文数据库，涵盖全世界最广泛的科技、医学和社会科学领域的科技文献，以及高品质的网络资源，2004年11月开始提供服务，2007年推出了系列特色服务，提供追踪、分析研究成果并将其可视化的智能工具[3-5]。

Web of Science（WoS，http://isiknowledge.com）由Thomson公司于1997年将SCI、SSCI、AHCI等数据库整合创建而成，2016年由科睿唯安公司收购[6]。WoS是一个基于Web整合构建的数字研究环境，通过强大的检索技术和基于内容的连接能力，将高质量的信息资源、独特的信息分析工具和专业的信息管理软件无缝地整合在一起，兼具知识检索、提取、分析、评价、管理

与发表等多项功能，从而扩展和加大了信息检索的广度与深度[7]。

CiteSeerX自动引文搜索引擎（http://citeseerx.ist.psu.edu/index）最早由NEC公司研制开发，公开在互联网上提供免费服务，被誉为全球最大的科学文献免费全文索引搜索引擎[8-9]。CiteSeerX的更新系统于2007年投入运行，新系统在可用性、全面性、及时性和成本效率等方面得到改进，科学文献传播和知识获取功能进一步增强[10]。

下文从上述数据库的文献收录范围、元数据描述、检索功能、统计分析功能4个方面进行比较分析。

2 来源文献收录范围与元数据描述比较分析

2.1 来源文献收录情况

在收录文献的学科范围方面，Scopus和WoS除了科技领域之外，还收录了生命科学和社会科学领域的文献；CiteSeerX主要关注计算机和信息科学的文献；DISC主要涵盖自然科学与工程技术领域的文献。

在地域语种方面，Scopus覆盖多语种，包括英语、法语、德语、日语、意大利语、俄语、西班牙语、汉语等；WoS以英语为主；CiteSeerX仅收录英语语种文献；DISC虽然也收录多语种文献，但并没有收录中文期刊。

在收录时间方面，Scopus和WoS的来源期刊可追溯到上百年前，CiteSeerX从1948年开始收录，而DISC来源期刊从2006年才开始收录。

在来源文献类型方面，Scopus和WoS收录多种类型的文献资源，包括期刊文献、会议论文、图书资源、专利资源等；CiteSeerX收录预印本、期刊、会议等；DISC只收录期刊文献，不包括学术会议、图书、专利与技术报告等。

在数据规模方面，截至2020年1月底，Scopus收录2.5万多种来源期刊，17亿条引文；WoS的SCI收录236个学科中超过1.1万多种世界上最具影响力的期刊，累计17亿条引文；DISC收录6 000余种来源期刊，1 100多万条来源文章，3.8亿条引文。

在更新频率上，Scopus每天更新，WoS每周更新，CiteSeerX实时更新，DISC则每周或更长时间更新。

DISC在学科范围、地域语种、文献类型、数据规模与更新频率等方面与其他数据库还存在很大差距。

2.2 元数据描述

元数据描述详尽与否体现了数据库对文献内容的揭示深度，直接影响到数据库的检索与分析服务效果。下文通过对相关字段的统计分析，分别对4个引文数据库的文献元数据、作者元数据、机构元数据和来源出版物元数据的描述内容进行比较（见表1）。

分析结果表明，Scopus和WoS不仅提供了十分丰富的元数据内容，并且提供了作者、归属机构方面的ResearcherID、ORCID等规范编码，为其开展丰富多彩的数据库应用奠定了良好基础。

与另外3个数据库相比，DISC的元数据描述内容相对较少，可供利用的元数据字段项较为有限，来源出版物元数据只提供了期刊名和ISSN或E-ISSN号，文献元数据描述内容也很不丰富，作者元数据和机构元数据的描述内容则更少，没有专门的元素集描述。

3 检索功能比较分析

在检索功能方面，Scopus提供了文献检索、作者检索、归属机构检索3个主要的检索入口，并在检索结果查看中提供了选择查看次要文献（参考文献）的功能，还可以对来源出版物进行检索；WoS提供了基础检索（文献检索）、引文检索、作者检索和化学结构检索等多个检索入口；CiteSeerX提供了文献检索、作者检索和表检索3个检索入口，同时可以在检索时选择是否包含引文；DISC提供NSTL所有文献的检索、引文库来源文献检索、引文检索3个入口，并可以对来源文献进行浏览和检索。

3.1 文献检索

4个引文数据库提供的文献检索功能和检索结果排序输出方式的数量见表2。

Scopus、WoS和DISC均提供了3种检索方式，DISC虽然提供了组合检索与高级检索选项，但可检索字段与结果筛选项，与Scopus、WoS相比还存在很大差距。

在检索结果输出方式上，4个数据库各有特色，Scopus提供了自定义输出字段；WoS可直接与写作工具相结合；CiteSeerX对单篇文章可以进行添加列表和添加标签；DISC提供了添保存检索历史的功能。

表1 元数据描述内容对比

项 目	Scopus	WoS	CiteSeerX	DISC
文献元数据	标题、文献号、DOI、CODEN、PubMedID、关键词、摘要、学科类别、语种、被引频次、领域加权引用影响、PlumX、参考文献、相关文献、文献类型，期刊（题名、ISSN、出版年、卷、期、页码范围），书籍（书籍、标题、丛书、卷号、期、页码范围、出版日期），会议（出版物名称、会议标题、会议地点、会议日期），专利（专利名称、发明人/申请人、年份）	标题、文献号、DOI、入藏号、IDS号、关键词、摘要、学科类别、语种、被引频次、参考文献、相关文献、文献类型，期刊（题名、ISSN、出版年、卷、期、页码范围、子辑、增刊、特刊），书籍（书籍标题、丛书、卷、期、页码范围、出版日期），会议（出版物名称、会议标题、会议地点、会议日期）	标题、关键词、摘要、引文上下文、被引频次、相关文献，期刊（题名、ISSN、出版年、卷、期、页码范围）	标题、关键词、摘要、语种、被引频次、参考文献、相关文献、文献类型，期刊（题名、ISSN、出版年、卷、期、页码范围）
作者元数据	作者姓名、ResearcherID、ORCID、其他姓名格式、地址、邮件地址、文献量、引文总量、被引文献、合著作者、合著文献、学科类别、h-Index、作者出版时间范围、参考文献	作者姓名、ResearcherID、ORCID、其他姓名格式、地址、邮件地址、团体作者、书籍作者、书籍团体作者、编者	作者姓名、地址、邮件地址、文献量	作者姓名、地址、邮件地址
机构元数据	机构名称、机构地址、名称变体、归属机构ID、出版文献总数、作者总数、专利数、合作机构、合著文献、按来源出版物的文献、按学科领域的文献		机构名称	机构名称、机构地址
来源出版物元数据	名称、收录年份、文献库订阅、ISSN或E-ISSN、类型、语种、学科范围、ISBN或E-ISBN，出版物影响（CiteScore等级、CiteScore百分比、CiteScore排名、SJR、SNIP）	名称、ISSN或E-ISSN、类型、语种、学科范围，出版物影响（当年影响因子、5年影响因子、JCR分区、JCR类别中排序），出版商（名称、地址、类别/分类）	类型、出版商（名称、地址）	名称、ISSN或E-ISSN

表2 文献检索功能对比

项 目		Scopus	WoS	CiteSeerX	DISC
检索方式	快速检索	√	√	√	√
	组合检索字段	21个	27个	8个	9个
	专业检索字段	64个	26个	-	9个
检索限制		3种	2种	1种	1种
检索结果筛选方式		10种	18种	2种	4种
检索结果排序方式		5种	7种	4种	5种
检索结果输出方式		6种	8种	2种	3种
检索结果输出格式		6种	4种	-	2种
结果查看	文献详情字段	18个	39个	6个	9个
	查看参考文献	√	√	√	√
	查看施引文献	√	√	√	√
	查看专利文献	√	-	-	-
全文获取	链接到出版商	√	√	-	√
	开放获取期刊	-	√	-	-
	免费下载全文	-	-	√	-

注：√代表对应数据库有此功能，-表示没有对应功能

在检索结果输出格式上，Scopus基本提供了目前主流文献分析工具所使用全部格式；WoS提供了4种输出格式；CiteSeerX没有提供检索结果批量导出的功能；DISC只提供文本和CSV两种格式。

在检索结果浏览和获取方面，Scopus的普通期刊文献详情页提供了18个字段的内容，可以查看参考文

献、施引文献，还可以查看专利检索结果；WoS则提供了高达39个字段的内容，获取全文时，还提示了哪些是开放获取期刊；CiteSeerX文献详情页面提供了6个字段的内容，并提供多个全文链接选项，可以免费下载全文；DISC提供了9个字段的内容，可链接到出版商数据库，并可下载全文，也可以通过NSTL进行原文请求。

3.2 作者检索

Scopus、WoS和CiteSeerX都提供了单独的作者检索入口，DISC未提供作者检索入口，但在组合检索中可用作者姓名进行检索。Scopus与WoS提供了6个不同的检索字段，CiteSeerX提供了作者姓名1个检索字段。

在作者检索结果显示方面，Scopus显示内容最全，有11个字段；WoS其次，有10个字段；CiteSeerX提供了3个字段的内容；DISC在组合检索中，用作者姓名检索结果与文献检索结果显示一致，提供了题名、作者、文献出处、被引频次、全文链接5个字段的内容。

3.3 归属机构检索

Scopus提供了专门的归属机构检索入口，可以通过机构名称检索某一机构的文献产出情况和了解机构的影响力，检索结果显示内容和排序文献都比较完善；WoS、CiteSeerX与DISC没有专门的归属机构检索入口，但在组合检索中提供了机构检索的相关字段，可以进行机构检索，检索结果显示相关机构发表的文献列表。其中WoS对机构检索的文献结果也与文献一样可进行多种维度的分组统计与排序，DISC的机构检索也只是在组合检索时，可以根据机构名称进行检索，检索结果显示与作者检索一致，提供了5个数据项。

3.4 来源出版物检索

Scopus和DISC提供了专门的来源出版物浏览和检索入口，Scopus检索功能和结果内容显示都比较完整。WoS和CiteSeerX没有专门的来源出版物检索入口，但在组合检索中提供了来源出版物检索的相关字段，可以进行检索，检索结果显示所检索出版物的文献列表。DISC的来源出版物提供了4项内容，可以对来源出版物的题名、ISSN、年份和卷期进行浏览和简单检索。

4 统计分析功能比较分析

对四大引文数据库的检索统计分析功能进行比较分析，结果见表3。

（1）检索结果分组统计与排名。对检索结果进行多维分组、统计与排名是文献计量分析的基本内容。在4个数据库中，WoS对所有检索入口所检到的文献结果均可进行16种分组统计与排名，并可以进行图表显示；CiteSeerX没有提供分组功能，只是对检索结果进行被引频次的排序；DISC提供的分组方式中，关键词云功能是其他3个数据库没有的。

（2）文献引文分析。作为引文数据库，文献引文分析功能是最重要的内容，而引文分析报告则是最好的呈现。Scopus和WoS都对检索到的文献提供了引文分析报告，包含多项影响力度量指标。DISC的引文分析功能包括被引量和年被引量两方面，另外可以查看施引文献，并提供文献引用提醒，显然与其他数据库相比还存在不小差距。

（3）作者与归属机构分析。4个数据库中，只有Scopus提供了完整的作者和归属机构分析功能，WoS虽然没有提供独立的分析入口，但是通过检索，可以获得比较全面的作者分析内容、机构的基本情况与科研产出情况，DISC可以查看作者合作网络。

（4）来源出版物分析。Scopus提供了完整的来源出版物详情，并设置多维评价指标，同时还可以通过图表对多种来源出版物进行指标的可视化对比分析；WoS提供的当年SCI期刊影响因子、5年平均影响因子、JCR类别、JCR类别中的排序和JCR分区等已经成为被广泛应用的、权威的文献计量指标；CiteSeerX和DISC没有提供来源出版物分析功能。

5 改进DISC建设与服务的建议

通过对4个引文数据库的对比分析可见，DISC与其他3个引文数据库，尤其是与Scopus和WoS相比，在来源期刊收录范围、对数据的描述及数据深度挖掘分析等方面都还存在不小的差距，系统目前提供的服务功能较为有限，需要在以下方面加以改进。

5.1 加强DISC的基础功能建设

现有的DISC引文数据库收录来源文献的学科范围

表3 检索统计分析功能对比

项 目	Scopus	WoS	CiteSeerX	DISC
检索结果分组统计与排名	出版年份、国家/地区、归属机构、作者、来源出版物、文献类型、学科类别	出版年份、国家/地区、归属机构、机构扩展、作者、来源出版物、文献类型、学科类别、基金资助机构、基金授权号、丛书名称、会议名称、编者、团体作者、语种、研究方向	被引次数	出版年份、作者、来源出版物、被引次数、关键词云
文献引文分析	引文总数、排除作者自引、排除书籍引用、年引文数、FWCI指标、引文基准分析、PlumX度量指标、h-index指数、引文提醒、可视化报告、引文耦合、共享作者、共享关键字	引文总数、排除作者自引、高被引文献数量、平均引用次数、引文提醒、可视化报告、施引文献、相关记录	引文总数、引文提醒、引文及上下文、引文耦合、共享关键字、同被引文献	引文总数、年引文数、引文提醒、施引文献
作者分析	姓名、所属机构、国家/地区、作者ID、ORCID、可能匹配的作者、发文量、出版物、类型、年份、学科类别、被引频次总计、年被引、h-index指数、合著作者、合著文献数量、可视化、出版时间范围、参考文献历史、出版物历史	姓名、研究领域、所属机构、ORCID、可能匹配的作者、发文量、论文组数量、出版物、类型、年份、学科类别、被引频次总计、去除自引、h-index指标、合著作者、高被引文献数量、平均引用次数、施引文献、去除自引的施引文献、可视化、出版时间范围	-	姓名、所属机构、发文量、出版物、年份、合著作者、合著文献数量、合作网络、可视化
归属机构分析	名称、名称变体、归属机构ID、地址、机构文献量、专利、作者量、学科领域占比、来源出版物占比、合作机构、合著文献量	名称、地址、机构文献量、研究领域、基金资助机构、被引频次总计、去除自引、高被引文献数量、平均引用次数、施引文献、去除自引的施引文献	-	-
来源出版物分析	出版物名称、ISSN、学科类别、类型、出版商、开放获取、文献量、CiteScore、SJR、SNIP、被引次数、未引用比例、评论文献百分比	文献量、当年SCI期刊影响因子、5年平均影响因子、JCR类别、JCR类别中的排序、JCR分区	-	-

不够广泛、语种不够丰富、文献类型较为单一、数据规模较小，在整体基础建设方面不够系统和完整。可以进一步拓展来源文献收录范围，丰富完善元数据描述内容，提高DISC数据库基础建设的系统性与完整性，为检索与引文分析功能提供更好的支撑。

5.2 丰富检索入口与检索结果展示功能

检索功能是评价数据库优劣的重要指标，而周到的检索功能可为用户提供更多的便捷，保证数据库的检索效率[11-13]。

通过对比可以发现，成熟完善的引文数据库提供了文献、引文、作者、机构、来源出版物等多个独立的检索入口，而在引文分析功能方面，对作者和归属机构的分析至关重要、必不可少，DISC应增添这方面的检索入口，提供更加丰富的检索字段，检索字段多可使数据库具有很强的引文统计分析功能和文献检索功能[14]，因此，DISC在组合检索方式中，还有很大改进余地。

DISC在检索结果筛选、检索结果排序、检索结果输出方式以及检索结果输出格式等检索结果多样化展示方式上明显偏弱，可选择性较小，应增加多途径分组与排序方式，如文献数量、归属机构、出版日期、被引频次、来源出版物、国家/地区、学科类别等，为用户使用数据库提供更多更好的使用体验。

5.3 拓展引文分析功能

对检索结果进行更多维度的统计与排名比较分析，会使得分析结果更加客观、准确，便于用户开展更加广泛的文献计量学方面的应用。

文献引文分析功能和评价指标是引文数据库的核心价值[15]。Scopus提供了多项影响力度量指标；WoS提供的引文报告，其内容包括多项指标参数，被学术界广泛认可和使用；CiteSeerX作为自动引文数据库，其提供的信息也颇具特色；而DISC的引文分析功能没有提供任何引文分析的报告和评价指标，非常有必要进一

步完善。

作者、归属机构和来源出版物分析功能方面，DISC也相当欠缺，没有相应的分析评价功能，需增加相应的评价指标与服务功能。

参考文献

[1] 国际科学引文数据库（DISC）系统简介 [EB/OL]. [2019-07-17]. http://disc.nstl.gov.cn/disc/view/m09/A0901.xhtml.

[2] 任慧玲，杨滨，黄利辉，等. NSTL国际科学引文数据库医学外文期刊引文数据加工流程和加工技术研究 [J]. 医学信息学杂志，2009，30（3）：19-21.

[3] Scopus [EB/OL]. [2020-05-10]. https://www.elsevier.com/solutions/scopus.

[4] 刘筱敏，孙媛，和婧. Scopus与SCI来源期刊影响力差异化分析 [J]. 中国科技期刊研究，2014，25（9）：1171-1177.

[5] 王继红，肖爱华，张贵芬，等. Scopus数据库研究综述 [J]. 中国科技期刊研究，2016（12）：1241-1247.

[6] Web of Science [EB/OL]. [2020-05-10]. http://isiknowledge.com.

[7] 杜永莉，陈锐. Web of Science最新版检索及其新功能介绍 [J]. 现代情报，2005（10）：110-111.

[8] CiteSeerX [EB/OL]. [2019-06-10]. http://citeseerx.ist.psu.edu/index.

[9] 马雪. CiteSeer~X—免费获取计算机类外文科技文献的搜索利器 [J]. 内蒙古科技与经济，2009（21）：131-132.

[10] 宋歌. 引文搜索引擎CiteSeer~x设计原理及检索 [J]. 中国索引，2008（3）：2-7.

[11] 王婧，华薇娜. 国内外文科引文索引数据库检索功能比较 [J]. 新世纪图馆，2011（1）：42-44，73.

[12] 高星. 国外四大引文数据库比较分析 [C] //中国医学科学院/北京协和医学院医学信息研究所/图书馆学术年会. 2009.

[13] 曹志梅，王凯. 我国四大引文数据库比较分析 [J]. 情报学报，2002（4）：481-485.

[14] 翟中文. 中美引文数据库比较研究 [J]. 图书馆工作与研究，2011（9）：71-74.

[15] 谢暄，蒋晓，王燕，等. Scopus与Web of Science比较分析 [J]. 科技与创新，2017（4）：8-10.

作者简介

杨增秀，女，1969年生，硕士，馆员，研究方向：数据库建设与数据管理、数字资源长期保存，E-mail：ziyang_1@163.com。
张桂玲，女，1976年生，硕士，高级工程师，研究方向：文献服务。
杨海超，男，1990年生，硕士研究生，工程师，研究方向：数据库建设与数据管理。
冯艳君，女，1978年生，本科，工程师，研究方向：数据库建设与数据管理。
赵卫华，女，1976年生，硕士，工程师，研究方向：情报研究。
张欣，女，1981年生，本科，工程师，研究方向：资源建设。

Viewing the Construction and Development of DISC from the Comparative Analysis of International Famous Citation Database

YANG ZengXiu ZHANG GuiLing YANG HaiChao FENG YanJun ZHAO WeiHua ZHANG Xin
(China Machinery Industry Information Institute, Beijing 100037, China)

Abstract: This article selects Scopus, Web of Science and CiteSeerX citation databases vs. DISC for comparative analysis. Compares and analyzes several dimensions of literature collection, metadata description, retrieval function and statistical analysis function, and summarizes the characteristics of the current international advanced citation database in terms of scope, search route, analysis and evaluation in order to improve the DISC furtherly. The relevant suggestions of DISC construction and service work were put forward based-on the above research.

Key words: Citation Databases; Scopus; Web of Science; CiteSeerX; DISC

（收稿日期：2020-04-10）