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

关于发现系统的问题与思考

□ 秦鸿 / 电子科技大学图书馆 成都 611731

摘要：文章基于三种国外主流的发现系统Summon、EDS和Primo，对图书馆界在发现系统评估与应用中出现的一些共性问题和分歧提出思考与分析，重点评述引进发现系统的必要性，评估发现系统的主要因素，以及展望未来发现系统的发展前景。

关键词：发现服务，发现系统，Summon，EDS，Primo
DOI：10.3772/j.issn.1673—2286.2012.07.002

2012年6月11日至13日，“2012高校图书馆发展论坛暨数字图书馆前沿问题高级研讨班”在苏州举行。会中，“发现系统”成为热点之一。

网络级发现服务系统是近年来国内外图书馆界非常关注的一个新应用，它采用预索引的元数据仓储技术，提供简单、单一的检索入口，将文献资源的一站式整合检索提升到一个全新的境界。自2009年Summon发布之后，发现系统发展很快，国外知名大学图书馆纷纷应用，国内的北大、清华等高校图书馆也在近期引进了不同的发现产品，起到了极大的带动效应。

然而，关于发现系统目前也存在着不少争议，会前会后，都有业界人士表达了不同的意见和观点。笔者曾对Summon、EDS、Primo三种国外主流的发现系统做过较为详细的研究与测评，并与商家和一些业内人士有过较为深入的沟通和交流，本文仅就发现系统应用领域中一些共性的问题和分歧提出自己的思考和分析，为发现系统的评估提供参考。

问题一：用户真的需要发现系统吗？

就笔者的观察，关于发现系统图书馆界存在着两种声音，一种声音是站在支持的立场，认为图书馆庞大的资源体系特别是数字资源体系必须要整合，发现系统是目前为止最佳的产品，引进发现系统非常必要；另一种声音是持怀疑的态度，认为引进发现系统是图书馆一厢情愿的做法，用户可能根本不需要。在苏州会议的头脑风暴环节，台上的专家之间也产生了分歧，约翰·霍普金斯大学的张东明老师说：“我们学校的研究人员觉得发现系统没什么用。”而新泽西大学的王永明老师说：“我不同意你的观点，发现系统在我们学校用得很好。”

就此问题笔者曾与西安交通大学的邵晶馆长交流，就邵馆长的观点来看，发现系统本来就是给普通用户而不是高端用户用的。换言之，它的受益面应该是广大的本科生和普通研究生，而教师可能只看本领域最顶尖的几种刊物就够了。研究生不论是开题还是写综述还是为导师的科研项目搜集资料，都是图书馆数字资源最大的用户群，而大部分研究生的信息素养远远低于我们想象的水平，所以他们真的需要一个简单易用的检索系统，这也是Google受到追捧的原因。从众多学校的使用效果来看，发现系统的确极大地提升了数字资源的使用量。

另一方面，不同学科的文献获取有不同的需求和特点，发现系统也并不是放之四海而皆准的普适工具。华中农业大学图书馆的潘宏指出：“生物和医学领域的研究人员需要跨库检索核酸序列、蛋白序列、蛋白结构这样大量的信息资源，所以他们首选的检索工具既不是SCI也不是Primo或Summon，而只会是PubMed，发现系统诞生之前是，之后也会是。”[1] 约翰·霍普金斯大学是一所以医学见长的高校，这会不会是他们的研究人员不认同发现系统的原因之一呢？同理，化学化工专业的师生检索文献时，他们会选择发现系统还是依旧

选择用SciFinder呢？发现系统对不同学科的适用度如何？这是值得我们去进一步探究的问题。

问题二：图书馆需要发现系统吗？

基于以下四个原因，图书馆需要一个发现系统：

（1）数字资源的资源量、经费和使用量已超过纸质资源；

（2）高端用户基本上不到实体图书馆来，数字资源是他们利用图书馆资源的主要方式；

（3）集成管理系统中旧的管理模式无法承担对数字资源的高效管理；

（4）除了发现系统，还没有一种理想的数字资源管理平台。

目前发现系统的发展非常快，因为价格的因素，我们可以暂时持观望的态度，但是不能忽视资源整合这一必然的大趋势。

数字图书馆已经发展那么多年了，但是图书馆基本上还是围着一个集成管理系统在运转，这合理吗？相比于纸质资源的加工，图书馆从事数字资源加工的人员少得可怜，数字资源也需要整序、揭示、评估和推广，试问全国那么多图书馆，有几个馆开展了数字资源编目？所以引进发现系统并不是一个简单的购买行为，而是管理重心的转变和业务流程的重组，发现系统的实施是一项系统工程，需要对本馆资源进行清晰的梳理，数字资源的状况随时都可能变化，需要有专门的数字资源管理人员进行后期的数据维护，需要确切了解用户需求以调整发现系统的配置，也需要和商家合作进行更深入的数据挖掘和分析。总之，图书馆不能是一个简单的Buyer，而应该是系统需求的制定者和产品发展的主导者。

问题三：评估发现系统什么因素最重要？

我在苏州会议的报告中列出了四大类20多个评估指标，四大类分别是元数据、架构与功能、检索与界面、商务因素。

元数据是发现系统的存在基础，对元数据的评估重点还不在其数量，虽然商家都极力宣传他们的发现系统的数据规模达到了多少亿的数量级，但是其中超过一半的数据是报纸或新闻类的非学术性数据，而且存在不少的重复数据。另外据测试，三种发现系统对主流全文数据库如ScienceDirect、Springer、Wiley、Talor&Francis、IEL等的覆盖度都比较好，均能满足绝大多数的用户需求，按照目前各产品迅猛的发展趋势，资源覆盖率的提升只是时间问题，所以评估元数据更应关注其质量。元数据质量分为数据深度和数据规范性两个层面，数据深度有“薄、厚”之分，薄元数据仅包含“题名、作者、来源”等少量字段，厚元数据则包含了主题、摘要、全文等更丰富的信息。数据规范性则决定了结果集是否能较好地进行归并和去重，比如作者是姓在前还是名在前，题名中的特殊符号是否完整，卷期号的标注方法是否一致等，稍有不同都会造成检索结果不能去重。实际上因为元数据来源复杂，在数据清洗时进行数据规范是非常困难的，我们的实际测试发现各个发现系统产品都存在不少的重复数据，可以说在去重方面都做得不够好。

架构与功能方面应关注系统的部署模式、与OPAC的整合度、可扩展性等指标。特别是与OPAC的整合度，标志着发现系统仅仅是一个以数字资源为主的整合系统还是图书馆所有资源的整合服务系统。最低层次的整合是对本馆纸质资源提供馆藏信息链接，即呈现馆藏位置和实时流通信息，目前三种国外产品都能做到这一点；再深一点的整合是支持用户互动，如为图书写评论、加个性化标签等；更深层次的整合是嵌入OPAC的所有功能，如预约、续借、个人图书馆的功能（借阅历史查询、电子书架等），使发现系统完全取代OPAC，真正实现图书馆所有资源的一站式服务。

作为一个整合检索系统，检索功能是否强大及其界面展示是否友好也是应该关注的重点，其中又以检索结果的相关度排序最为重要，因为越排在前面的检索结果越可能被查看。从刘颉颃等在广州大学城开展的一项关于发现系统的调研可见，检索结果的相关度排序被用户认为是最有用的功能[2]。Lynda Duke也在ALA2012年会的“发现和检索工具比较”专场报告中汇报了由两所学校的90余名学生参与的一项测试，对5个检索系统的检索效率进行对比（EDS、Summon、Google Scholar、2个传统的图书馆系统），结果同样支持了相关度排序是至关重要的评价因素，甚至92%的参与学生都只在第一页找他们需要的文章，而几乎不使用“分面”功能去筛选[3]。

商务因素也是决定一个产品能否被引进的重要因素。其中又以价格因素最为关键。目前国外的发现产品价格普遍还比较昂贵，除了首次实施费用外，年度费用

也不菲，使很多图书馆望而却步。窃以为，价格是目前影响发现系统普及的最主要因素。另外，如果引进国外产品，是否能提供本地支持也非常重要，包括本地服务支持和本地技术支持。

总之，目前的三种国外产品各有千秋，Summon和EDS都是数字资源提供商出身，以内容见长，他们的发现系统还主要定位于数字资源的整合，同时兼容了本馆纸质资源的一部分揭示功能，而Primo以系统集成商起家，以技术取胜，在深度整合OPAC方面略胜一筹。具体到一个图书馆，在选择发现系统方面要综合考虑本馆的资源状况和经费情况，选择最适合本馆的发现产品和服务。

问题四：中文资源的整合如何解决？

目前，国外的三种主流发现系统对中文数字资源的覆盖度都不够好，原因是中文数据库商还不太愿意将元数据提供给其他发现产品开发商，另外，某些数据库商出于保护自有发现服务系统核心竞争力的原因，也不可能开放其元数据。发现系统中不能涵括中文资源，这对实现资源一站式检索的目标是一个重大的制约。众所周知，即使在国内的研究型大学中，中文资源的下载量也是外文资源下载量的数倍，尤其是如果发现系统的主要用户群定位在中低端用户，那么这些用户的主要资源需求一定是中文资源，发现系统不能有效整合中文资源是很多图书馆暂时还不愿考虑引进发现系统的主要因素之一。

所幸这一状况正在得到改善，国外发现产品提供商与中文数据库商家之间的洽谈一直在紧锣密鼓地进行中，目前，维普公司已与三家国外产品提供商进行了签约，向其提供维普期刊数据库的元数据，在Primo系统中已灌装了维普的元数据，大约有3000多万条。此次苏州会议上，某发现产品商宣布其与方正Apabi也达成了正式合作意向，据悉，另一国内知名数据库商与国外发现产品的合作也有重大进展，未来我们可以期待，国外的发现产品在整合中文资源方面会做得更好。当然我们也更期待国内的发现产品，如超星百链、万方学术搜索和CALIS的E读等，能迎头赶上，打造中国自己的发现系统品牌。

问题五：发现系统看起来还是一个整合检索系统，何谓“发现”？

关于这一类的质疑不在少数，很多馆员在了解了“发现系统”这个新名词后，都发出类似感慨，甚至有人说“发现系统看起来和联邦检索也没什么不同”，言下之意，发现系统的功能被夸大了。

那么如何理解发现系统的“发现”作用呢？

首先，发现系统脱胎于早期的联邦检索系统，虽然在界面上仍然比较相似，但系统的原理与架构已完全不同了。联邦检索系统是一种分而治之的异构检索，检索效果依赖于各个数据库系统的自有功能，在检索速度、检索结果的去重和排序等方面存在难以克服的缺陷，而发现系统预建了统一的元数据仓，是一种同构检索，保证了秒级的响应速度和有效的排序算法。而且，联邦检索系统只能整合本馆资源，而发现系统的一个最大特点是不仅能整合馆内资源，而且能整合馆外资源，包括大量的OA资源。我们知道，没有被揭示出来的资源等于不存在，因为用户无法利用到它。发现系统检索出的馆外资源虽然不能直接获取全文，但让用户知道有这样一篇文献存在，这就是“发现”的意义之一，在馆际互借与文献传递渠道畅通的情况下，从发现到获取也只是一步之遥。

其次，目前的发现系统还有一些功能是帮助用户发现有用资源的，如Primo的bX学术推荐服务和bX Hot Articles热点文章推荐服务，前者利用数据挖掘分析文献之间的关联，将全球其他研究者也关注过的相关论文无缝地推送到读者面前，弥补了个人依靠检索词搜索进行资源发现的不足，后者侧重于提供某个主题最近几个月内的热门文章，以体现某个领域的研究趋势，这两项服务使“搜索”更接近“发现”。

然而，这些发现功能是否足够？距离我们理想的“发现”境界是否还相差甚远？

张甲老师说：“信息在读者动态使用中形成智慧集聚，读者第一线的知识总是高于图书馆或系统所能想象的范围。目前的发现系统虽然模仿了Google的一个检索框，却没有抓住读者点击进入后的知识过滤行为的特点和共性。知识探索工具如果不是以读者为第一线的设计，我们就会在图书馆的框框里跳不出来。”他还以Google的知识图谱为例强调了知识关联的重要性[4]。

诚然，不同的读者类群对知识的过滤行为差异甚大，发现系统应该足够智能地进行用户行为分析。比如，目前的发现系统在进行相关性排序时主要还是看检索结果命中的字段和主题，只有Primo在其相关性排序的专利技术ScholarRank中考虑了用户的身份，能够

结合用户的专业、学历等信息来判断检索结果的相关性，并根据是新检索还是缩小或扩大检索范围来动态调整结果的显示顺序。这可以视为用户行为分析的一种尝试，但要真正做到智能化的搜索，还要依赖关联数据和语义搜索技术的发展。

关联数据让数据开放并连接在一起，万维网发明者Tim Berners-Lee爵士说："关联数据是和万维网的发明一样的巨大变革，我们正在通过万维网从文件互联网走向数据互联网。"他认为关联数据就是一箱箱数据，当通过开放标准关联在一起时，从中可以萌发出很多新事物[5]。王波说："在5年之内，图书馆检索技术的进步将主要体现为关联数据的应用，就是扩大图书馆管理系统对数据的关联、分析、联想和推送能力。"[6] 林海青说："关联数据的功能体现在两个方面，一是数据整合，即通过关联数据将各种数据源无缝地关联起来，成为一个广域分布的数据库。二是数据发现或挖掘，关联数据对关系形式化描述，形成一张关系地图，使得机器可以通过理解和处理数据之间的各种关系，发现新的数据。[7]"

而所谓语义搜索，是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正意图，并以此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果[8]。

如果按照关联数据和语义搜索的框架来设计和打造发现系统，发现系统就是一个开放的而非封闭的，智能的而非机械的，真正能为不同用户提供知识发现的理想产品。

问题六：发现系统是不是只是一个过渡产品？

在苏州会议的头脑风暴环节中，清华大学的陈武说"发现系统还只是一个过渡产品"，得到了许多与会者的认同。笔者对上一个问题的思考似乎也印证了发现系统目前还不太成熟。然而软件产品的发展从来都不可能一步到位，从现在的眼光看，当年的联邦检索系统不也是过渡产品吗？Web2.0号称永远的Beta版，就是在不断创新中臻于完善。

世界已经进入了大数据时代，全球知名咨询公司麦肯锡在研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素[9]。数据就是知识，对图书馆而言，以云服务为基础的发现服务系统是数据的大规模集聚与定制化分布的结合，发现系统的海量元数据以及以发现系统为起点的大量的用户检索和使用数据就是一种大数据，对这些数据的去冗分类、去粗取精、深度挖掘和关联分析将为图书馆带来难以想象的价值，发现系统对未来图书馆资源利用和管理模式产生的深层次影响才刚刚开始。

参考文献

[1] 苏州数图会后的学习[EB/OL]. [2012-06-17]. http://librarysalon.com/space-1453-do-blog-id-11537.html.
[2] 刘颉颃,陈定权,郭蝉.用户对图书馆资源发现系统功能的期望——基于广州大学城高校图书馆学生用户的调研[J].图书情报工作,2012(7):27-31.
[3] ASHER A D, DUKE L M, WILSON S. Paths of Discovery: Comparing the Search effectiveness of EBSCO Discovery Service, Summon, Google Scholar, and Conventional Library Resources [EB/OL]. [2012-06-30]. http://crl.acrl.org/content/early/2012/05/07/crl-374.full.pdf+html.
[4] 读panhong的"苏州数图会后的学习"[EB/OL]. [2012-06-18]. http://librarysalon.com/space-21-do-blog-id-11550.html.
[5] 怒放的关联数据:你为什么应该关注?[EB/OL]. [2012-06-30]. http://www.pbdigg.net/show.php?tid=24461.
[6] 王波.图书馆正进化为超智能[J].大学生,2012(5)(上).
[7] 关联数据和资源导航[EB/OL]. [2011-06-27]. http://blog.sina.com.cn/s/blog_4c725fcc0100vz55.html.
[8] 语义搜索及框计算:从百度查生僻字谈起[EB/OL]. [2010-03-04]. http://www.cnr.cn/allnews/201003/t20100304_506100493.html.
[9] 大数据(百度百科)[EB/OL]. [2012-06-30]. http://baike.baidu.com/view/6954399.htm.

作者简介

秦鸿（1972-），女，副研究馆员，研究方向：数字图书馆的理论与技术。E-mail: qinh@uestc.edu.cn

Problems and Thoughts about Discovery Systems

Qin Hong / Llibrary, University of Electronic Science and Technology of China, Chengdu, 611731

Abstract: Based on evaluation and application of three kinds of foreign mainstream discovery systems, Summon, EDS and Primo, this article puts forward thought and analysis about some common problems and disagreements. The necessity of discovery system construction, key factors of discovery system evaluation, and development prospect of discovery system are mainly discussed.

Keywords: Discovery service, Discovery system, Summon, EDS, Primo

（收稿日期：2012-07-02）