·综　述·

国外电子资源在线使用统计研究述评*

索传军　王建朋
(郑州大学信息管理系　郑州　450001)

〔摘　要〕本文总结分析了国外电子资源在线使用统计研究的主要内容,指出该领域的研究重点是要解决数据的规范化和可获取性的问题,研究的难点在于获取深层次数据的方法和数据管理工具的开发,并对重要研究成果 COUNTER 规范进行了评析。

〔关键词〕电子资源　在线使用　统计

1　电子资源在线使用统计研究的源起

20 世纪 90 年代以来,电子资源在图书馆预算中所占的比重越来越大。以 ARL 成员馆为例,从 1992 ~ 1993 与 2000 ~ 2001 年的成本数据可以看出:ARL 成员馆在电子资源方面的投入从 3. 6% 升到 16. 25%,几乎是 1992 ~ 1993 的五倍[1]。在这种情况下,有关电子资源的选择、管理、服务、效能等问题就逐渐突显出来。

电子资源在线使用统计,概括地说,就是对利用一定的方法和手段对电子资源的使用情况进行科学的统计。具体地说,就是某个单位或组织,利用一定的方法和手段对某种电子资源在某段时间内的用户访问、检索和下载数据等情况进行科学的统计。

电子资源的使用统计数据是电子资源服务绩效评估的重要依据。通过使用数据的分析与研究[2],图书馆可以了解本馆订购的电子资源的使用情况,并以此作为制定与调整电子资源建设政策、争取资源建设经费、开展资源推广与用户培训工作的可靠依据。因此,对电子资源在线使用数据的收集、传递、管理、分析的研究就深入开展起来。纵观国外的研究,我们发现,其研究内容主要集中在以下三方面:①使用统计的标准化研究,②使用数据获取方法的研究,③数据收集分析系统的研究。

2　使用统计的标准化研究

2.1　有关的主要研究项目

统计的标准化研究包括数据元素和定义的标准化、数据收集和处理原则的标准化、数据传递格式的标准化。解决好标准化和规范化的问题,不同来源的数据才具有可比性,不同的电子资源之间才能进行绩效大小的比较,使用统计才具有真正的意义。为此,国际组织和标准化机构进行了深入的研究并制定了相关标准。其中最显著的有:

(1)图书馆联盟国际联盟基于网络的信息资源使用统计指南(ICOLC Guideline for Statistical Measures of Usage of Web - based Information Resources)。在这个指南中,ICOLC 制定了一套网络的信息资源的使用统计指标,1998 年发布了第一个版本,2001 年又发布了修订版,详细说明了使用统计数据收集和报告的一系列最低要求,同时还就数据的保密性、访问统计数据的权限、统计报告的格式和传递方式提出了明确要求[3]。

(2)ARL E - Metrices 项目。项目的研究任务由佛罗里达州立大学的信息使用管理及政策研究所(Information Use Management and Policy Institute)来承担,主要研究人员为 Charles. R. McClure、francis Eppes、Wonsik"Jeff"shim、John Carlo Bertot,参与单位为 ARL 的部分成员馆,项目分三个阶段进行。项目制定了统计数据收集的规则、一套推荐使用的统计数据、一份数据收集处理指南,并且对数据的可获取性、有效性进行了实地验证[4]。

(3)ISO2789 信息与文献国际图书馆统计数据标准(2003)。由于电子资源及电子服务在图书馆

*本文系国家自然科学基金项目"电子资源在线使用统计与绩效评估"(项目编号:70573099)阶段性研究成果。

中所占的比例越来越大,ISO 的图书馆统计数据标准 ISO2789(Information and Documentation – International Library Statistics)在 2003 年的修订版中增加了一个附录 A:电子图书馆服务使用评价。标准指出了需要收集的数据元素,并对每个统计指标进行了定义,说明了它们是如何收集的。

(4)NISO 图书馆统计标准 Z39.7 – 2002。NISO Z39.7 全称为信息服务和使用:图书馆和信息服务机构统计指标 – 数据字典(Information service and use: Metrics & statistics for libraries and information providers – data dictionary),是由美国国家标准组织(ANSI)制定的关于图书馆统计方面的国家标准。标准于 1968 年制定,在 2002 年的修订版中加入了数字馆藏使用和服务的统计指标[5]。

(5)COUNTER 项目[6]。COUNTER 是 Counter Online Usage of Networked Electronic Resources 的缩写。它的目标是研制一系列普遍接受的、国际化的实施规范,用于管理不同种类的电子资源的使用数据。目前 COUNTER 已发布了四大类电子资源(电子期刊、数据库、电子图书、参考工具书)的使用统计规范,主要包括在线使用数据的定义、数据收集报告的原则以及数据传递的格式等内容。COUNTER 规范是目前电子资源在线使用统计研究取得的最重要的成果。因此,我们在本文第 5 部分重点对 COUNTER 规范做专门的介绍和分析。

2.2 国外对使用统计标准化的研究的明显特征

(1)权威组织积极参与。如 ISO、NISO 等国际和国家标准化权威组织都开展了调查研究,制定了相关标准,并不断吸收其它组织和机构的研究成果对标准进行完善。它们的参与影响和带动了其它组织和机构的研究工作,有力的推动了标准化的进程。

(2)不同组织之间互相借鉴研究成果。如 COUNTER 规范在制定数据的定义时,除自己定义以外,尽量采用 ISO、NISO、ICOLC 等已有的定义,借鉴它们的研究成果,解决仍然存在的问题。正因为如此,COUNTER 规范得到了标准化组织、图书情报界和数据商的广泛认可和支持,成为电子资源使用统计事实上的标准。

(3)重视实证研究和用户培训。在 ARL 的 E – Metrics 项目和 IMLS 资助的公共图书馆统计和绩效测度项目中,先把准备收集的数据在图书馆和数据商中进行讨论,听取它们对数据元素的看法,了解哪些数据的可获取性差,同时对这些数据的收集原则和方法进行探讨,最终形成了一套广泛接受的核心统计指标。它们还编写了统计手册,指导用户进行使用数据的收集。这些工作增加了用户对核心数据的认同,对使用统计的标准化非常有利。

3 使用数据获取方法的研究

使用数据的获取方法有两种:一是本地日志分析,二是从数据商获取数据。由于从数据商获取数据只是"被动"的接收数据的过程,不存在数据的析出问题。因此,图书情报界更多的是对日志分析法的探讨和研究。具体包括以下几个方面:

(1)日志文件的格式和能从中获取的数据。文献[7]介绍了一般日志文件格式(the common file format),并举例说明了参与日志文件(referrer log file)、代理日志文件(agent log file)包括的数据。最后,归纳了能从中获取的数据:请求的资源、用户的 IP 地址、日期和时间、请求是否成功等。文献[8]、[9]、[10]指出通过处理日志能获取:用户请求的数量(hits)、文件的字节数、不同格式文件的数据(如 HTML 文件)、每个 IP 的请求数、每个文件的请求数等。

(2)日志文件分析的局限性。文献[7]分析了动态 IP 和缓存(caching)对数据获取和准确性的影响。文献[8]指出,我们不能准确获取的三类数据:日志中没有记录的某一具体类数量(如用户的数量)、日志中记录但不完整的数据(如用户请求的数量)、不合理的推断(如 hits 和 use 是不相等的)。文献[11]分析了动态 IP 和缓存(caching)等多种影响日志分析的因素,指出为什么网络统计没有意义。

(3)日志文件分析软件的介绍、分析和选择。文献[7]列出了免费的软件(analog、wwwstat 等)和商业软件(WebTends Log Analyzer、Faststat 等),并介绍了它们的特点和适合分析的日志格式。文献[8]、[10]还归纳了日志分析软件的特点:界面友好性强、输出文件格式灵活多样(Html、Word、Text 等)、支持多种日志文件格式、实时分析功能等。

(4)日志文件分析应注意的问题。文献[10]指出,在我们进行日志分析获取数据时应注意:理解数据的确切含义、选择合适的分析软件、和服务器管理员合作、加强日志分析方面的培训、日志文件的传递和保存等。

另外,文献[12]还对两种数据获取的方法进行了比较分析,说明了用本地获取的数据来检测数据商的数据的可行性。并且指出由于图书馆经费、技术等因素,在今后很长时间内图书馆不得不依靠数据商来获取详细的使用数据。

这些研究介绍了日志分析的基本原理,分析了能从中获取的数据和它的局限性。但由于图书馆资

金、技术等方面的原因,对日志的分析还不深入,只是能够获取一些最基本的数据,对数据挖掘的研究更少,不能获取深层次的使用数据。因此,图书馆更多的是依靠数据商来获取详细的数据。但这些数据仍然存在不真实、不完整、不及时的问题。总体来看,国外对数据收集方面的研究还不系统、不深入,特别是在数据挖掘、数据商数据的验证和获取个性化数据方面还需进一步分析。

4　数据收集分析系统的研究

由于不同的数据商以不同的格式(online, e - mail, text, excel, csv, etc)和方式(statistic web site, e - mail)提供使用报告,而且记录不同的使用数据元素(hit, searchs, sessions, type, etc),数据的收集费时费力,分析的结果也不准确。因此,开发一种工具来解决这些问题成为研究的焦点,最有影响的是 DLF 的 ERMI 项目、NISO 的 SUSHI 项目、ERUS 项目。

(1)数据图书馆联盟(DLF)的 ERMI(Electronic Resource Management initiative)项目[13]。这个项目旨在形成一个综合的、基于生命周期的电子资源管理系统,来管理电子资源的费用、版权、使用等信息,为图书馆提供一个有效的电子资源管理工具。ERMI 于 2001 年起动,2002 年 5 月和 NISO 一起组成了项目指导组,并组织图书馆员、系统开发商和相关组织代表形成了咨询委员会。项目形成了一系列文档详细定义了系统的功能需要,并帮助形成数据标准。

(2) SUSHI(standardized usage statistics harvesting initiative)项目[14]。现在遵从 COUNTER 报告的数据商以 EXCEL 表单的形式或容易转换成 EXCEL 的格式分别传递使用报告给用户。用户从多个数据商获取并转换成 EXCEL 格式的管理成本非常高。然而,还没有一个标准的数据池(standard data container)来自动的合并、存放这些数据。SUSHI 的研究就是要设计一个基于 web 的、自动化的请求/响应协议,用来下载 xml 格式的使用数据并存放在数据池中,从而形成一个标准的计算机对计算机(machine to machine)的数据获取模型。从而解决电子资源管理系统的数据获取问题。

(3) ERUS(E - Resource Usage Statistics)项目[15]。它开始于 2003 年,由 Simmons Coolege, Trinity College 和 Villanova University 合作研究。目的是设计一个综合的、基于 web 访问的数据库,用来收集、报告、分析来自不同数据商的数据。经过调查和电子资源分类之后,2004 年完成了数据库的初步设计,并设计了 web 界面,ERUS 系统模型基本形成。由于电子资源类型和数据商性质的多样性,系统开发的最大问题在于标准化数据的自动化获取。因此,ERUS 项目正吸收 ERMI 和 SUSHI 的研究成果来推动系统的进展。

以上述三个项目为代表,国外在这方面的研究针对性强,重点解决标准数据的自动获取问题;实践性强,注重图书馆、数据商、系统开发商的合作;有用性强,系统的设计充分考虑与图书馆其它决策系统的结合,为图书馆提供一个综合的解决方案。因此,系统的开发进展顺利,成效显著。

从对研究内容的分析中我们发现,国外对电子资源在线使用统计的研究关键是要解决两个方面的问题,即数据的可获取性和规范性问题,而研究的重点和难点在于获取深层次数据的方法和数据管理工具的开发。

5　COUNTER 规范评析

COUNTER 规范是目前电子资源在线使用统计研究取得的最重要的成果。得到了美国出版协会、国家信息标准化组织等 11 个机构的大力支持。越来越多的数据商开始遵从 COUNTER 规范。是否遵从 COUNTER 规范已经成为图书馆订购电子资源时首先关注的问题。因此,我们重点对 COUNTER 规范进行介绍和分析。

5.1　COUNTER 已取得的成果

COUNTER 项目于 2003 年 1 月发布了第一版实施规范,2005 年 4 月发布了第二版,这两版都是针对电子期刊和数据库类的电子资源。经过调研 2006 年 3 月又定义了电子图书(e - book)和参考工具(e - reference)类电子资源的统计规范。

COUNTER 通过规范数据定义、数据收集和处理的原则、数据的传递为使用统计提供了一个标准框架,为解决数据的不可比性提供了一个有效途径。它要求遵从的数据商必须提供符合 COUNTER 规范的使用统计报告,并规范了使用数据的定义,有助于理解和分析各种数据和信息;它还说明了数据采集和处理的原则,如规定在 HTML 格式的链接上间隔不足 10 秒的双击只被记为一次请求,在 PDF 格式的链接上间隔不足 30 秒的双击只被记为一次请求,这些都保证了能够收集到标准的数据;另外,COUNTER 还要求使用报告的传递必须是 EXCEL 格式或者是容易转换的格式。要上载到一个以密码控制的网站上,用户可以用密码随时获取,当有数据更新时要通过邮件通知用户。至少每月提供一次报告。要保留上年度和本年度至今的数据。

5.2　目前的研究工作

COUNTER 项目仍在顺利进行中,当前的研究

主要集中在三个方面:一是研究和探讨更高级别的使用报告和选择性报告,给用户提供更多的可用数据。二是研究和定义其它类型电子资源的统计规范,扩大 COUNTER 规范的应用范围,充分发挥 CONTER 规范的实用性。三是吸收用户对 XML DTD 的反馈,研究和发展基于 XML 的数据传输协议。

5.3 现阶段存在的问题

COUNTER 规范得到各方面的认同和支持,但现阶段它又有自身的局限性,表现在:

(1)定义的统计指标过少。电子资源使用统计的目的是为绩效的测度提供依据,但 COUNTER 规范仅仅定义了极少数的统计指标,不能满足绩效评估对深层次使用数据的需要。

(2)数据商不完全遵从 COUNTER 规范,仍存在数据的不可比性。现在遵从 COUNTER 规范的数据商越来越多,但仍占总数的一小部分。图书馆收集到的使用数据由于没有遵从统一的标准,因此不能很好的结合和比较。COUNTER 规范的重要作用不能很好的发挥出来。

(3)COUNTER 规范不强制数据商遵从基于 XML 的数据传递协议。COUNTER 规范最近几年定义和发展了 XML DTD,但并不要求数据商一定遵从。这给用户自动化获取、传递、处理、分析数据造成了困难,增加了图书馆的管理成本。

总之,从以上对国外电子资源在线使用统计研究的分析可以看出,国外对该问题的研究给予了高度重视,围绕关键问题开展了深入的理论和实践研究,取得了显著的成就。同时在标准化、规范化的进程中还存在一些有待解决的问题。由于电子馆藏发展阶段的不平衡,我国图书馆界已经开始重视电子资源的使用统计和绩效分析问题。因此,我们有必要借鉴国外的研究方法和成果,积极参与国际标准的探讨和国内标准的制定,分析研究中外文电子资源使用数据的获取方法,在此基础上开发适合我们自己的电子资源管理系统。

(来稿时间:2006 年 3 月)

参考文献:

1. Julia C. B. measures for Electronic Use: The ARL E - Metric project. (2006 - 1 - 5). http://www. arl. org/stats/newmeas/emetrics/Blixrud_IFLA. pdf

2. 郭依群. COUNTER——网络化电子资源使用统计的新标准. 大学图书馆学报,2005(2):20 ~ 23. Resources. (2006 - 3 - 5). http://www. library. yale. edu/consortia/2001webstats. htm.

4. ARL E - Metrices. (2006 - 1 - 15). http://www. arl. org/stats/

5. Z39.7 Information service and use: Metrics & statistics for libraries and information providers - data dictionary. (2005 - 11 - 25). http://niso. org/emetrics/current/complete. html

6. COUNTER. (2005 - 11 - 25). http://www. projectcounter. org

7. Kathleen Bauer. Who Goes There? Measuring Library Web Site Usage. (2005 - 11 - 26). http://www. infotoday. com/Online/OL2000/bauerl. html

9. Joel Riphagen, Alaina Kanfer, Ph. d. In Search of the Elusive User: Gathering Information on Web Server Access. (2006 - 1 - 15). http://www. ncsa. uiuc. edu/

10. John Carlo Bertot, Charles R. McClure, William E. Moen, Jeffrey Ru bin. Web Usage Statistics: Measurement Issues and Analytical Techniques. Government Information Quarterly, 14(4):373 ~ 395

11. Goldberg, Jeff. Why web usage statistics are (worse than) meaningless. (2005 - 12 - 26). http://www. cranfield. ac. uk/docs/stats/

12. Duy and Liwen Vaughan. Usage Data for Electronic Resources: A Comparison between Locally Collected and Vendor - Provided Statistics. The Journal of Academic Librarianship, 29(1):6 ~ 22

13. DLF Electronic Resource Management Initiative. (2006 - 3 - 25). http://www. diglib. org/standards/dlf - erm02. htm

14. NISO standardized usage statisties harvesting initiative. (2006 - 3 - 16). http://www. niso. org/committees/SUSHI/SUSHI. comm. html

15. ERUS - Electrotic Resource Usage Statistics. (2006 - 3 - 18). http://www. simmons. edu/ ~ andersoc/erus

A Review on Electroinc Resource Online Usage Statistics

Suo Chuanjun Wang Jiaanpeng (Department of Information mamagement)

〔**Abstract**〕 This paper analyses the main content of electronic resource online usage statistics study and indicates that the study is to solve two key problems: the standardization and the availability. The difficulties are methods of obainting usage statistics and development of usage statistics management tools. It finally analyses limitations of the COUNTER Code of Practice.

〔作者简介〕 索传军,郑州大学信息管理系教授、博士、系主任,主持和参加国家级项目研究 8 项。发表核心期刊论文 40 余篇、出版学术著作 3 部。王建朋,郑州大学信息管理系研究生。