图书馆数字资源访问统计研究

马建霞
中科院资源环境科学信息中心 兰州 730000

摘要 本文分析了图书馆数字资源访问统计的重要性，国内外图书馆数字资源访问统计的研究现状，探讨了进行图书馆数字资源访问统计的方法，比较了数字资源访问统计的指标，并就目前我国图书馆数字资源访问统计存在的问题进行了讨论。

关键词 数字图书馆 数字资源 数字化馆藏 使用评价
分类号 G253

Usage Evaluation of Library Digital Collections

Ma Jianxia
Scientific Information Center for Resources and Environment Seciences

Abstract

Digital collections represent a significant and growing part of the academic library's collection. Measuring usage of library digital collections is complex, however, and the lack of standard metrics makes it especially difficult to develop a framework for evaluation. This article explores the benefit of usage evaluation of library digital collection, reviews the work and research having been carried out. And discusses the ways to get the data of usage of library digital collection, the indexes of usage evaluation of library digital collection, Limitations and the questions of usage evaluation of library digital collection are also presented.

Keyword:Digital Libraries；Digital Library Collections；Digital Resources；Electronic Resources；Usage Evaluation；Use Evaluation.

图书馆数字资源是图书馆馆藏资源中以数字形式保存的和借助于计算机网络可以利用的信息资源的集合。[1]图书馆数字资源从来源上主要包括购入式数字资源（主要是指由出版商或数据库商生产发行的、商业化的正式出版物，有数据库、全文电子期刊和电子图书等，其中数据库又包括参考数据库(书目、文摘、索引)、全文数据库和事实数据库）、自建式数字资源（自建数据库，指图书馆根据各馆的用户需求和特色，建设的专题数据库)、开发式数字资源（网络信息导航)。

[1] 索传军.论述字馆藏的质量评价.中国图书馆学报,2004,30(152):43-46

1 图书馆数字资源访问统计的重要性

随着网络和计算机技术的发展和我国数字图书馆建设的展开，数字资源逐渐成为图书馆信息资源建设的重要组成部分，据美国研究图书馆协会统计，数字资源的费用占其文献费用的 20%，近年来我国各类图书馆在数字资源建设方面的投入也呈现上升趋势，一些科学图书馆也基本达到数字资源的费用占其文献费用的 20%左右的水平。

图书馆数字资源的访问统计对于以用户为中心的数字图书馆建设和服务模式而言，尤为重要。

1．1．1 有利于信息资源建设和开发的科学决策，使图书馆有限的经费得到有效的应用

通过对数字资源的访问统计，可以了解图书馆数字资源的使用情况，经过与数字资源的成本分析比较，对图书馆信息资源建设和开发的科学决策提供参考依据，使得图书馆有限的资源得到最有效的应用。

1．1．2 有利于针对用户需求，促进数字资源的整体优化

由于图书馆 20%的资源满足了用户 80%的需求，图书馆可以根据数字资源的访问情况的统计分析，促进数字资源的整体优化建设：调整数字资源的学科分布，调整参考数据库、全文数据库、事实数据库、电子期刊、电子图书、自建数据库、网络信息导航的比例，使其结构逐步优化，更大程度地符合用户需要，从而调整采访策略，改善图书馆资源结构。

1．1．3 有利于了解用户的访问行为，改善图书馆服务

通过对数字资源结构的调整和使用情况的分析，了解图书馆在资源宣传、服务中存在的问题，促进检索服务、咨询服务、培训服务的开展，使电子资源及其服务更符合用户需要，从而提高其利用率，降低成本。

通过了解用户的访问行为，可以为图书馆网站的信息构建（IA）提供参考，提高图书馆网站的易用性。

1．1．4 有利于图书馆绩效评价的完成

近年来，为了加强对各个图书馆绩效的评价，高校图书馆系统、科学图书馆系统和公共图书馆系统都展开了相应的评估工作。各个图书馆要客观地完成评估表格，非常需要对图书馆数字资源的使用情况进行统计，从而提供真实客观的评估依据。

从图书馆数字资源的主要来源看，通过研究图书馆数字资源的使用情况，数字资源供应商可以根据资源的使用情况，提供针对用户需求的产品，制定恰当的价格模型，根据用户的访问行为，改进数字资源产品的检索功能和网站功能，得到更好的收益和回报。

2 图书馆数字资源访问统计的现状

目前，国外对图书馆数字资源的访问统计的研究与工作已经展开。

在研究计划方面，美国和欧盟针对图书馆数字资源的访问统计已经展开了一些针对性的研究计划，比如，由美国研究图书馆协会资助的 E-Metric[2]项目、由美国多个机构(包括 ARL、JISC、NISO 等)资助的 COUNTER[3]项目、欧盟 Telematics for Libraries Programme 支持的 EQUINOX[4] 项目等，这些项目多为研究制定描述电子信息服务和资源的统计指标和绩效测度及其方法。

在相关的标准方面，面对新的信息环境和图书馆形态，一些组织开始尝试将新的电子资源绩效评估标准融入原有相关标准／指南的框架。例如 NISO 在 2004 年批准了图书馆和信息提供者信息服务和利用的测度和统计数据字典[5]（NISO Z39.7-2004 Information Services and Use: Metrics & statistics for libraries and information providers--Data Dictionary），该标准在传统图书馆工作的基础上，还特别增加了网络服务、网络资源、网络运行的新的测度方法，这套数据字典将逐渐纳入美国图书馆统计工作，成为美国图书馆统计工作的参考依据。

ICOLC[6]1998年制定的《网上索引、文摘和全文资源使用统计测度指南》(Guidelines for

---

[2] http://www.arl.org/stats/newmeas/emetrics/index.html

[3] http://www.projectcounter.org/index.html

[4] http://equinox.dcu.ie/

[5] http://www.niso.org/emetrics/index.cfm

[6] ICOLC. GUIDELINES FOR STATISTICAL MEASURES OF USAGE OF WEB-BASED INFORMATION

Statistical Measures of Usage of Web—Based Indexed,Abstracted and Full Text Resources)提供了一套网络化信息资源使用的绩效测度指南。2001年的修订版明确了网络信息使用数据统计的最基本要求，并提供在隐私、保密、获取、传递和报告形式方面的指导。

ISO ISO/CD 11620[7]也在传统服务统计指标的基础上，结合 ICOLC 和 COUNTER 的研究，进行了图书馆数字资源测度及其定义、方法的描述。

国内随着公共图书馆、大学图书馆、科学图书馆系统图书馆评估工作的进行，图书馆界开始逐步重视对图书馆数字馆藏、图书馆数字化信息服务的评估。

文献 2 中提出了数字资源后评估的概念，但是对图书馆数字资源访问统计等后评估的方法和指标体系尚未全面展开评论。一些图书馆自行开发了基于 jsp 或者 asp 的图书馆网站访问统计软件，一些数字图书馆系统，如清华同方的ＴＰＩ、北京拓尔思的ＴＲＳ、浙江天宇的ＣＧＲＳ等等也提供了相应的统计功能，但是尚没有一款商业化的软件针对图书馆的各种类型的数字资源提供一揽子的访问统计方案。

目前，多数图书馆对于购买的资源的用户访问统计，依赖资源提供商提供的统计数据，对于自建的数字资源，以及图书馆网站的访问统计，往往是自行开发。存在的问题是尚无完备的数字资源访问统计的标准，急需访问统计的一致性科学性和连续性，以便为图书馆的决策提供可信的决策支持。

3 图书馆数字资源访问统计的方式

web 服务器在工作时，时刻将 WWW 访问的结果记录在一些 log（日志）文件中，通过对服务器日志的分析可以得到以下信息:

(1) 通过对访问时间进行统计,可以得到服务器在某些时段的访问情况;

(2) 对访问者的ＩＰ进行统计,从中可以判断主要是那些用户在访问 Web 服务器;

(3) 对访问请求的错误进行统计和分析,可以找出有问题的页面加以改正;

(4) 对访问者请求的ＵＲＬ进行统计,就可以判断出读者对那些页面的内容最感兴趣,对哪些页面的内容不感兴趣。

各种web服务器日志文件的格式和内容大致相同。根据W3C的标准[2]，一般Web日志都包括诸如用户的IP地址、请求时间、方法(GET／POST等)、被请求网页或文件的URL、发送／接收字节数、协议版本等信息。表1列出了几种不同类型的Web日志。

RESOURCES <http://www.library.yale.edu/consortia/2001webstats.htm

[7] http://www.libraryjournal.com/article/CA411564?display=FeaturesNews&industry

服务器	Web 日志
Microsoft IIS	2002-10-29 17:45:37 10.100.183.33C-GET /localstart.asp -3426 638 HTTP/1.0
Apache	202.106.175.93 - - [03/Apr/2002: 10:30:17 + 0800]”GET/INDEX.HTML HTTP/1.1”2000 419
IBM WebSphere	129.42.19.99 -[07/Feb/2002:4:41:55 -0600]”GET cgi-bin/commerce3/EecMacro/orderdspc.d2w/report/HTTP/1.0”200 4957

表1 Web 日志数据[8]

但这些日志文件信息一条一条的数目很大，用户难以直接从 log 文件获得直观的结果。对日志文件的分析，可以借助一些商业性的或者源代码开放的软件完成。其中比较好的开放源代码的日志分析软件有：AWStats、webalizer 等。

从日志文件提供的信息进行统计和分析,就可以对整个网站有一个数字化、精确的认识,从而对网站的设计和内容改善和调整,使图书馆网站更好地为读者提供服务。

数据库的使用情况属于后评估指标，主要用于更新、续订数据库时使用，一般在图书馆购买资源提供商的数字资源时，应该要求由出版商或数据库商提供使用报告，再据此进行各类分析。

目前出版商 / 数据库商提供的统计报告常用的相关统计指标有：

① 检索次数(searfh / query)：用户在某一个数据库中提出检索式的次数.

② 登录次数(session / sign on)：用户打开某个数据库的次数。

③ 下载文摘 / 全文(abstract / fulltext page/image)：用户在某一个数据库中下载到本地客户机中的文摘或全文篇数。

代理服务器(Proxy Server)是一种服务器软件，它的主要功能有：设置用户验证和记帐功能，可按用户进行记帐，没有登记的用户无权通过代理服务器访问 Internet 网，可以对用户的访问时间、访问地点、信息流量进行统计。

目前代理服务器软件产品十分成熟，功能也很强大，可供选择的服务器软件很多。主要的服务器软件有 WinGate 公司的 WinGate Pro、微软公司的 Microsoft Proxy、Netscape 的 Netscape Proxy、Sybergen Netwo rks 公司的 SyGate 等，这些代理软件不仅可以为局域网内

[8]张川,肖金升,周振,胡运发. 具有访问时间完整性的 web 日志方法. 计算机应用与软件.2004(2):105-107

的 PC 机提供代理服务，还可以为基于 Novell 网络的用户，甚至 UNIX 的用户提供代理服务。目前绝大部分 Internet 的应用都可以通过代理方式实现。大多数代理服务器软件产品具有登记内部网用户访问外部网的日志记录，有些产品还可以直接将日志记录到数据库中。根据日志记录文件或数据库，可以统计内部网每个用户的网络流量以及上网时间，甚至可以按服务网络类型（如：HTTP、SMTP、FTP 等）分别进行统计。

通过web服务器的日志可以获得用户访问图书馆网站信息的情况，但是，这种方式需要对日志的格式进行了解，然后用相应的工具软件或者进行一定的开发来完成。还有一种获取网站访问情况的方法是利用asp或者jsp等网络脚本语言，利用它们内置的server、session、request对象等获取相关的信息，获取数据进行统计。比如：[9]利用Jsp我们可以用Jsp的内置request对象的获取参数方法request.getParameter("userid")，取到用户名；用request.getRemoteAddr()获取访问者的IP地址；通过request.getHeader("User-Agent")获取包含浏览器和操作系统的信息，然后用字符串分割substring()方法来分别得到浏览器和操作系统；通过Jsp的内置对象session的方法session.getCreation-Time()返回Session被创建的时间，而session.getLastAccessedTime()则返回当前Session对象最后被客户发送的时间，两者之差为停留时间。

主要分以下几个开发步骤：

（1） 确定将要统计的信息；

（2） 建立数据库；

（3） 实时的访问信息纪录，记录每次点击的信息，包括页面信息，用户信息，访问IP，访问时间；

（4） 实时信息的分类存储

（5） 显示方式的选择。可以用 Windows 的表格系统，也可以自行编制表格显示

利用这种方法相对比较简单，但是可获得的统计指标也有限。

除了上述几种统计方式外，还有基于路由器的流量统计、基于防火墙的流量统计、基于以太网广播特性的流量统计。但是这些方法之提供简单的流量统计功能，不能完全满足图书馆数字资源访问统计的目标。

---

[9]梁玉环，李村合，索红光. 基于 JSP 的网站访问统计系统的设计与实现. 计算机应用研究.2004(4):166-167

图书馆数字资源访问统计的指标

国际图书馆联盟认为，信息资源提供商对他们提供的特定的电子信息资源所提供的统计数据应该满足以下的最低需求。

1. 必须提供的数据元素是：

a) 会话（session）数量（或者登陆数量）number of sessions。为了满足政府机构和专业组织的报告的需要，应该提供会话数量或者登陆数量。在没有国界的网络环境中，会话数量的统计是一个粗糙的指标。

b) 提问数（number of queries），即经过分类的提问数量。一次检索是一次独立的知识查询。典型地，一次检索被记录为向服务器提交的一个检索表单，之后的浏览行为或者选定一个单独条目的行为没有表现为额外的检索，除非通过提交二次检索。立即进行重复的检索、双击或者其他用户的无意识行为都不应计入其内。

c) 菜单的选择数number of menu selections 如果数据的显示需要通过使用菜单来进行浏览，则应该提供这个指标（如一个电子期刊网站提供的基于音序和主体的菜单选择）

d) 全文的数量（打开的、下载的或者提供给用户的全文，这些全文都是由服务器控制的而不是由浏览器控制的）

期刊文章—按照期刊名称列出刊名和issn

电子书——按照书名列出书名和isbn

参考资料——按照改资源的内容单元（如字典的定义、百科全书的文章、传记等）

非文本型资源——按照自愿的文献类型(如图像、音频、视频等)

上述的每个数据元素应该按照每个特定的数据库提供商、按照每一组机构的IP地址或其他特别的元素（如账号），以及机构名称、按照协会名称、按照时间跨度（每月或者每年）分组描述，供应商还应该提供每天每小时的统计数据，而且还应该可以动态的集成几个月或者某一段时间的数据，而不用限制是当年数据还是由供应商限定的时间段。

为了了解图书馆数字资源的使用情况，确定数字资源的花费是否合理，ARL 的 E-Metrics 项目推荐的指标如下：

（1） 用户可检索的电子资源。包括：R1 电子全文期刊种数、R2 电子参考资源种数、R3 电子书的种数。

（2） 对网络资源和服务的使用情况。包括：U1 电子参考事务的数量、U2 登录电子数据库的数量（会话 session 数）、U3 电子数据库的提问和检索数量、U4 电子数据库的请求条数、U5 对图书馆网站和书目的远程访问次数。

（3） 网络资源和相关设备的花费。包括：C1 全文电子期刊的成本、C2 电子参考资源的成本、C3 电子书的成本、C4 图书馆对书目设备、网络环境等相关设备的花、C5 对书目设备、网络环境等相关设备的外部花费。

（4） 图书馆数字化活动。包括：D1 数字馆藏的大小、D2 数字馆藏的使用、D3 数字馆藏建设和管理的成本。

E-Metrics 的统计指标，既考虑了数字资源和数字化服务的访问量，还考虑了数字资源及其支持成本，便于从成本/效益的角度进行分析。

对于图书馆数字资源访问统计的指标，在我们常见的统计分析工作中，统计指标围绕什么被使用？谁在使用？如何使用？什么时候使用？为什么使用？哪些资料经常被下载？哪些资料被检索最频繁？资料检索来自哪些单位？哪个单位使用量最多等问题，通常采用数字资源提供商提供的访问统计数据与对图书馆网站及自建数字资源的访问统计相结合的方式，除了资源提供商提供的数据外，往往采用网站访问流量、访问者的 IP、网站点击次数、数字资源的点击次数、下载的篇数等指标。

与国外相比，我国图书馆的数字资源访问统计指标设定相对比较粗略，没有统一的、针对各种类型数字资源的一致的标准，而且统计指标往往仅仅反映了访问情况，未能与数字资源的购买和管理成本挂钩进行成本/效益分析。

5 图书馆数字资源访问统计存在的问题

随着各个图书馆在数字资源建设方面的积累和发展，图书馆数字资源的来源多样，既有通过远程镜像或者资源提供商服务器访问的数据，也有在本地镜像的数据，还有图书馆自建

的数字资源。尤其对于资料库不在馆内的情况，需要厂商配合协助，但是最大的问题在于没有办法从厂商那里得到充分的数据，或是厂商提供的数据不标准，或是提供的资料不是图书馆想要的，而且由于统计数据是由资源提供商提供，其客观性和真实性的保障机制弱。这样，正确及时的统计数据不易取得，

由于资源来源多样，统计指标不规范，不同的系统提供的统计报告五花八门，没有统一指标。统计指标定义混乱，不明确，例如“search 在大多数系统内被定义为用户发送检索式的次数，但有些数据库却用“query”来表示同样含义的指标，而 CSA 数据库则同时使用了“search 和“query”，二者的含义和区别并不明确。没有一致、标准、科学的统计指标体系，对用户访问统计的分析及其对图书馆决策的支持可信度就会降低。同时对于数字资源的访问统计指标还应该结合每种数字资源的类型、考虑数字资源服务学科的研究人员规模等参数。

图书馆数字资源的访问统计，是图书馆数字资源后评估的方法之一，目前的图书馆数字资源的访问统计存在统计指标不一致、不标准的问题，而且网站访问统计不能确定是否与使用者的目的相符，无法完全反映使用者真正的使用状况，因而，图书馆数字资源的后评估可以结合数字资源的访问统计、用户使用调查、用户访谈等方式完成。

图书馆数字资源访问统计的数据主要来自web server 的log files,目前法律上并无相关条文规定log file 资料的处理，但由于其中包含使用者的IP地址，应该与图书馆的流通记录一样，加以保密。不论图书馆决定如何分析log file 的数据，对于收集何种数据、谁能判读数据以及如何使用数据等，都应有详细的规定和说明，以免一时大意触犯了个人隐私权。未经个人用户同意，不能收集用户的个人信息，也不能将所收集的统计信息用于分析和识别用户个人信息。如果为提供特定服务必须采集用户的个人信息，必须向用户告知他的权利、个人信息用途及其保护方式，只有在用户知情同意的情况下才能基于该服务明确相关的个人信息。并且必须对合法采集的用户个人信息必须进行安全保管，未经用户同意不得公开，不得讲个人信息转给第三方，而且服务中止后，必须立即删除。[12]

6. 结论

图书馆数字资源访问统计做为图书馆数字资源后评估的方式之一，可以对图书馆、资源

[12] 张晓林、宛玲、徐引篪、宋小冬、王欣.国家科学数字图书馆数字资源采购的技术要求.中国图书馆学报.2004（7），14-19

提供商的决策提供有益的参考，但是目前尚未被图书馆决策者所充分重视。迫切需要做的是确定科学的、一致的、标准的评价指标体系，在此基础上，结合多种形式的数字资源的后评估，进行全面、客观、深入的统计分析，同时对使用统计的分析应该结合数字资源的成本进行成本/效益的考察。总之，对于图书馆数字资源的后评估，应该是数字图书馆时代，图书馆决策者合理有效地利用有限的经费，进行科学决策的参考。

参考文献

1索传军.论述字馆藏的质量评价.中国图书馆学报,2004,30(152):43-46

2肖珑、张宇红.电子资源评价指标体系的建立初探.大学图书馆学报,2002,(3):35-42

3 http://www.arl.org/stats/newmeas/emetrics/index.html

4 http://www.arl.org/stats/newmeas/emetrics/index.html

5 http://www.projectcounter.org/index.html

6 http://equinox.dcu.ie/

7http://www.niso.org/emetrics/index.cfm

8ICOLC. GUIDELINES FOR STATISTICAL MEASURES OF USAGE OF WEB-BASED INFORMATION RESOURCES <http://www.library.yale.edu/consortia/2001webstats.htm

9http://www.libraryjournal.com/article/CA411564?display=FeaturesNews&industry

10张川,肖金升,周振,胡运发. 具有访问时间完整性的web日志方法. 计算机应用与软件.2004(2):105-107

11梁玉环，李村合，索红光. 基于JSP的网站访问统计系统的设计与实现. 计算机应用研究.2004(4):166-167

12[同8]

13詹丽萍E-metrics 在数位图书馆使用评估的应用.http://p105.lib.nctu.edu.tw/2001conference/pdf/1-1.pdf

14 张晓林、宛玲、徐引篪、宋小冬、王欣.国家科学数字图书馆数字资源采购的技术要求.中国图书馆学报.2004（7），14-19

作者简介：马建霞 女，中科院资源环境科学信息中心副研究馆员，兰州大学图书情报学专业本科毕业，中科院文献情报中心硕士毕业，主要研究领域：信息资源管理，发表论文十余篇。