

古腾堡项目——数字图书馆的先驱者

□ 刘燕权 / 美国南康涅狄格州立大学 美国康州纽黑文 06515

刘晓东 / 北京师范大学管理学院 北京 100875

摘要: 古腾堡项目 (Project Gutenberg) 被普遍认定是最早出现的数字图书馆。1971年7月4日, 创始人 Michael Hart 将美国《独立宣言》输入计算机, 开启了古腾堡项目的第一步。40年后, 在 Mr. Hard 的领导和无数志愿者的帮助下, 该项目的免费电子书等目前已经超过4万本, 而其附属网站收集的电子书更是超过了10万本。古腾堡项目作为最早的数字图书馆在数字图书馆发展史中扮演了十分重要的角色。文章对该项目的历史背景、信息资源组织、服务和技术特点等进行客观的介绍和评价。

关键词: 数字图书馆, 古腾堡项目, 数字化藏书, 免费电子藏书

DOI: 10.3772/j.issn.1673—2286.2013.04.008

1 项目概述

1.1 背景介绍

古腾堡项目 (Project Gutenberg, 以下简称PG, 网址 www.gutenberg.org) 是一个非盈利的数字图书馆, 它免费为大众提供该馆的数字资源。1971年, 还在大学就读的 Michael Hart 利用做计算机实验的时间将《独立宣言》输入电脑 (Hart, 1992), 后又相继完成了美国《人权法案》、美国《宪章》以及《圣经》和莎士比亚戏剧 (Lebert, 2008) 等。这些数字资源被放到了当时处于起步阶段的因特网上, 成为一个可供大众免费使用的数据库。随着因特网在20世纪九十年代中期的迅速发展, 古腾堡项目所收藏的书籍数量也迅速增长, 截至2008年, 该项目的电子书籍一直以每年上千本的数量增加。Hart 致力于“推动电子书的创作和传播, 通过志愿者的努力完成文化内容书籍的数字化和收藏”, 他积极鼓励志愿者将他们认为有价值的书籍进行数字化。Hart 在2005年接受采访时强调, PG 对大型数字化项目如 Google Books 等的出现起到了重要作用 (Vaknin, 2005), 并指出该项目是网络信息扩散和传播的重要组成部分。

1.2 项目使命

PG 以“推动电子书的创作和传播, 用尽可能多的格式和语言向全世界提供数字资源, 冲破无知和愚昧的限制”作为自己的使命 (Hart, 1992), 其目标用户为任何需要信息资源的个人。Hart 表示“PG 的理念在于将信息、书籍和其他资源以电脑、程序和用户可以接受的电子数字格式提供给大众” (Project Gutenberg Literary Archive Foundation, 2010)。而要实现这一目标, 文本文档和开放格式是最佳选择。为此 PG 收藏的书籍多使用可以被近99%的人接受的“Plain Vanilla ASCII”格式 (American Standard Code for Information Interchange, 美国信息交流标准编码), 用户可按照自己的喜好对完成下载的文献进行装帧和格式调整。

1.3 合作机构

PG 数字图书馆是在个人和机构的共同努力下建立的。PG 文献档案基金会 (Project Gutenberg Literary Archival Foundation, 简称 PGLAF) 是一个非盈利机构, 为 PG 提供法律咨询, 监督 PG 运营并代表其接受捐赠, 组织筹资活动等, 其成立目的在于收藏文献和其他

智力作品；古腾堡项目合作中心（Gutenberg Consortia Center，简称PGCC）成立于1997年，致力于收藏、组织和传播在线数字资源。PG仅有少数兼职领薪的工作人员，主要依靠志愿者的合作实现书籍数字化、校正和其他工作（“About,” 2008）。另有分布式校对员机构（Distributed Proofreaders）负责监督校对志愿者提交到PG的数字资源。

PG在独立制作电子书籍的同时，也积极与其他机构合作生成多种类型数字资源：如PG与Librivox.org合作，用英语、西班牙语和中文在内的多种语言录制有声读物（MP3/VoggOrbis格式）（Librivox.org, 2012）；PG也与ClassicalArchives.com合作对公共领域的乐谱进行数字化。Carnegie Mellon University帮助PG优化了在线目录，大幅减少了浏览和使用的难度。IBiblio是PG主要的电子书发布网站，它管理PG的网页内容，提供多种辅助服务（Project Gutenberg, 2011c），并在现阶段与University of North Carolina – Chapel Hill共同管理PG。作为历史上第一个数字图书馆，PG也激发了其他国际性姊妹项目的出现，如欧洲、澳洲、加拿大项目等（Lebert, 2008; Project Gutenberg, 2011c）。PG与这些坚持相同理念的独立机构合作共赢，提升其在信息时代的影响力。

1.4 资金支持

作为一个非盈利机构，PG主要经济来源是社会捐

赠。Apple、IBM和Microsoft等公司都曾资助过PG的运营（Weller, 2000）。为了购买更多书籍完成电子化，PG鼓励用户使用多种途径实施捐赠，如使用PayPal、Flattr客户端或者直接使用支票和现金等^①。PG将其运作费用有意识地控制在较低水平，以确保其独立性，避免来自政治和经济利益方面的压力。PG对一般读者提供免费服务，但要求任何以盈利为目的而使用PG版权资源的组织和个人必须支付收入总额的20%作为回报。

2 资源组织

PG数字图书馆网站创建于1992年。

2.1 主页设计

PG网站设计简洁清晰，以黑色和栗色作为全站的主色调。网站主页包括欢迎内容及说明内容。网站标识以哥特式字体撰写，位于主页左上角的显著位置，下面分列不同的链接。用户可以通过输入标题、作者和主题检索书目和网页内容获取资源，还可以通过文献标题、作者和主题列表浏览文献。网页还提供捐赠入口和多语言选择链接，方便用户浏览和使用。页面中心展示了主要资源分类、新增项目及相关链接，包括自助出版入口、项目创始人Michael Hart的讣告、使用条款和网站地图等。

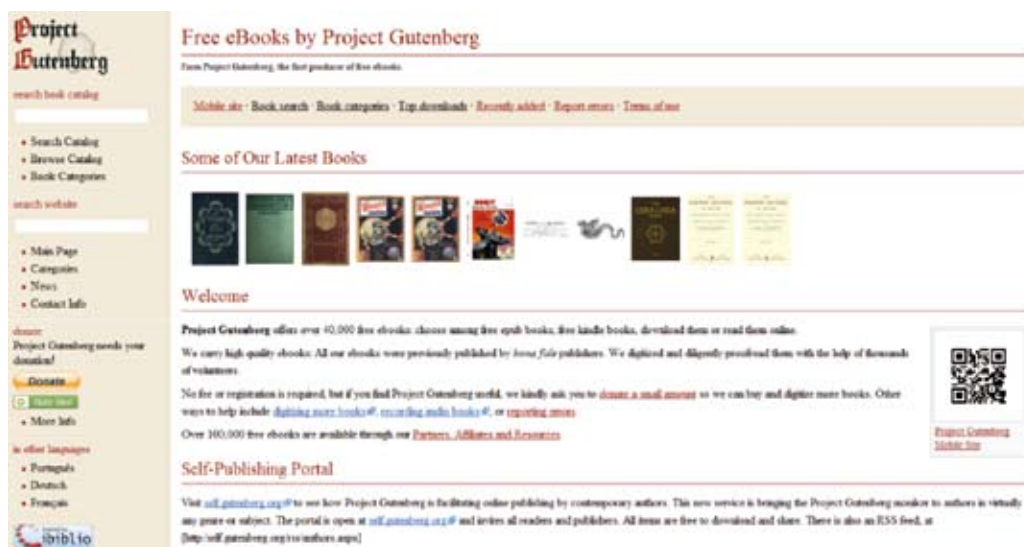


图1 古腾堡项目主页

^① http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Needs_Your_Donation捐赠页面，2012-10-11

2.2 数字资源

至今为止,已有超过五万名志愿者参与了PG的建设工作,使PG馆藏电子书超过42000部。PG的文本资源并非直接储存在PG数据库,而是通过FTP从相关网站上提取,PG提供索引指向特定文献(“Readers”, 2010)。馆藏中近29000部文献为英文文种,另有法语、中文、德语、西班牙语、葡萄牙语等外文文献。文献类型除了小说、诗歌、短篇小说、戏剧、食谱、参考资源和期刊文章等文本内容,还有非文本资源,如有声读物、乐谱、动态及静态图片等。同时,PG允许其他图书馆将其馆藏书籍的标题上传至本地目录,大大减少了查找数字资源的难度。

2.3 馆藏政策

PG并无正式的馆藏收集政策,Hart在其项目历史和理念(The History and Philosophy of Project Gutenberg)中提到:“PG基于实用原则选择电子书……我们选择大部分用户需要并会频繁使用的书籍”(Hart, 1992)。在收集图书过程中,PG严格遵守版权法,只发布进入公共领域的内容或者获得作者授权的书籍。PG向用户提供了精简版的版权法来判断所推荐的书目是否已进入公共领域。尽管重重预防,PG却依旧难逃版权诉讼,如Poul Anderson的The Escape则属越权文件,PG不得不将其移出馆藏。

2.4 数字化方法

从1971年到1999年,PG收藏的所有文档都是手动完成输入的。现阶段,项目雇用了大量不定时志愿者借助图像扫描仪和OCR(光学字符识别系统)软件等现代化工具完成数字化。潜在志愿者首先向网站提交他们认为有价值的数字化文献样本,经PG工作人员审阅确定其版权保护内容后即可开始进行数字化工作。经过审核的书籍诸页会被扫描成图像文件,并通过OCR软件将图像处理为文本格式。PG并不只是简单地提供文献书籍,而且非常注重产品的质量,因此需要校对人员对文本进行二次审核。每一部书在上传之前都需要经历三轮校对,三轮格式调整,并完成发布流程和检测用户友好性的流畅阅读环节。

2.5 元数据方法

PG使用美国国会图书馆的标准书籍元数据组,将作者、标题、简介和关键词制作为文献索引。通过作者、标题、编码和电子书序列描述文本文档(Project Gutenberg, n.d.; Project Gutenberg, 2006a)。Newby表示,PG使用EPUB Maker软件将文本文档转变为EPUB格式,该软件能在新书发布时从语法上分析元数据,并激发其他多种自动生成的格式。PG使用电子书序列号表示数据相关性,使用机读目录记录贡献者,都柏林核心集则被用于搜索引擎和目录软件(“Metadata,” 2010)。虽然PG声称其检索功能是“自产自营”的,但它却是由Yahoo!和Anacleto生成索引,并提供收藏书籍的离线目录。PG还有一个总索引文档罗列了所有馆藏目录,此文档包括对PG的描述、图书馆藏的完整列表和PG的正式执照。

2.6 文献组织

PG的资源按内容类型进行组织:通俗文学(如彼得潘)、经典著作(如圣经)和参考文献(如罗格同义词词典)。书籍也按照文献格式进行组织,如有声读物、音乐数据、乐谱、动态图片和静态图片等(Project Gutenberg, 2010b)。馆藏按照23个子目录编排,以字母顺排列为农业、教育、艺术、历史等多个部分,点击进入后,页面显示国会图书馆分类号标识,如教育类目为“L”,下设子类别教育史-LA、教育理论与实践-LB、教科书-LT等。收藏文献还按主题和主题词组织为“书架(Bookshelves)”,以字顺排列为动物、圣经、儿童小说、教育等共133个子类别,点击进入后,用户将看到该类目下的相关书目。

2.7 馆藏记录——描述和组织

一旦用户选定特定馆藏进行浏览,他们将进入该馆藏的详细记录。该记录包括三个标签:下载、书目记录、二维码。下载标签罗列了一系列可供下载的格式,如HTML、EPUB、Kindle、Plucker、QiOO Mobile和文本文档等,用户可选择所需格式从镜像网站快速下载。书目记录罗列了作者、标题、语言、国会图书馆分类号、主题、类别、电子书序列号、发布日期、版权状况和



图2 选定书籍描述页面

下载量。用户可通过“相似馆藏”浏览“其他用户也下载了……”的内容或同一主题文献。如George Herbert Betts的*The mind and its education*与*The teaching of history*相互关联。二维码界面直接将用户带入手机访问界面，方便用户直接将文献下载至手机阅读。用户还可以直接通过Facebook、Twitter和Google+分享检索结果。

3 信息服务

PG致力于为一般普通人群提供信息服务，故其网页结构简单，层次较少，设计单调无动画，用户界面智能性有限。但PG能够提供多种文献格式满足不同信息需求，如Librivox志愿者利用PG的内容制作免费的有声读物为残障人群提供服务，Kindle和iBook商店都链接至PG文档（PG, n.d.a）为普通信息用户提供通道等。需要指出的是，PG的主要目标在于实现信息可用性，致力于奉献而非出借文献，因此并不需要图书馆导向的服务内容如馆际互借、馆藏更新和数字参考服务等。

3.1 检索和浏览

在PG文献首页，用户可以直接浏览最新发布书籍。PG在线图书目录支持用户通过标题、作者、语言和最新发布进行浏览，作者和标题按首字母顺序排列，提供中文、法语、德语、拉丁语等多种语言内容。用户可根据资源类型浏览有声读物、编译文献、乐谱、动静

态图片等，或根据发布日期选择可能感兴趣的文献。PG还在页面上多处提示读者内容的更新情况。PG提供简单检索和高级检索服务，使用户可以直接在分类目录，如最受欢迎书目、最新发布内容、随意浏览和关注更新等分类目录中检索文献，也可设定多项条件完成检索，包括作者、标题、主题、语言、分类、国会图书馆分类号、文献类型和电子书序列号等（Project Gutenberg, 2010a）。除了网站自身的检索功能，用户还可以通过Anacleto、Yahoo!和Google等搜索引擎实现检索。Anacleto是一个每周更新的全文元数据搜索引擎，“它可实现对PG内容的字段检索和全文数据库检索”（Newby, 2007）。Yahoo!将PG纳入了其内容采编计划，可实现对书籍元数据的检索。Google可实现对于文章前100kb内容的检索，除了支持标准元数据（如标题、作者、主题等）进行检索外，还可使用替代标题、语言、文献类型、编码类型、创作时间、内容梗概和描述等内容检索。

3.2 联系信息

用户可以通过电子邮件、普通信件直接联系PG的工作人员。PG推荐使用电子邮件并提供了一系列邮件地址负责不同工作内容，如help2010@pglaf.org为用户提供普通服务，承诺两日内回复书籍和网页请求；纠错邮箱专供用户纠正书籍、目录、网站等出现的错误；copyright2010@pglaf.org提供版权内容咨询；用户还可以直接联系项目主管Dr. Gregory Newby获取深入信息^②。

^② http://www.gutenberg.org/wiki/Gutenberg:Contact_Information

3.3 移动服务

PG鼓励用户使用电子阅读器或移动设备阅读文献。用户可通过扫描二维码进入网站的移动版,每一项馆藏都有特定二维码。该移动版本提供检索功能,用户可通过作者、标题等检索馆藏并依据首字母顺序、受欢迎程度和发布日期对检索结果进行重新排序。作者使用关键词“Jane”检索,共得到69个结果,其中按作者名检索结果为59,按标题名检索结果为10,其中Jane Austen的*Pride and Prejudice*最受欢迎(下载量为18019次),Edith Bancroft的*Jane Allen: Center*为最新发布(2012-10-1)。用户还可对网页内容进行浏览,如最受欢迎作品、最新作品、随意浏览和新书发布等。点击进入后则显示按不同标准排列的书籍列表,如Beowulf为最受欢迎的书籍,下载量已达42341次,点击

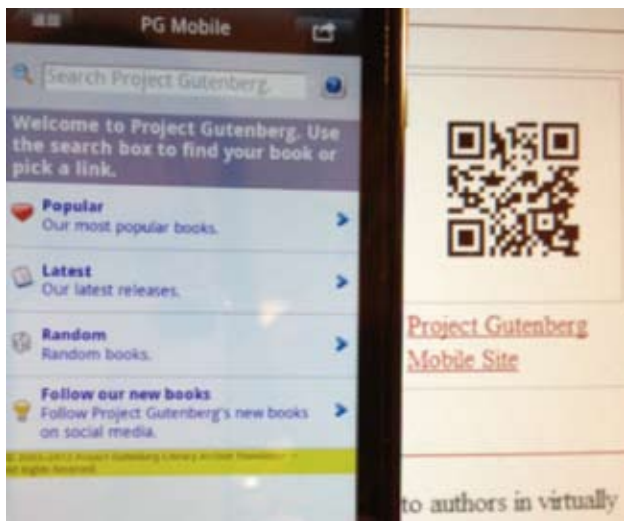


图3 PG的手机版页面

特定书目,则进入书籍描述页面,如作者、标题、电子书序号、语言、主题、分类和版权等内容,网页还提供多种下载格式(HTML、EPUB、Kindle、Plucker、Qioo Mobile等)和相关书籍内容,用户可浏览其他用户下载内容、同作者作品和同主题作品等。

3.4 Web 2.0工具: Blog和Wiki

PG网站也融合了一些Web 2.0技术,以增加网站的友好性。Newby指出,大部分非内容网页都是维基页面,这一灵活格式使PG工作人员和志愿者更便捷地增加或

修改内容。PG网站也拥有一个博客界面——古腾堡项目新闻(www.gutenbergnews.org),对PG的发展历史、创始人情况、特殊馆藏、合作机构等内容作了细致介绍,并提供标签云导航和社交网站链接。PG设有良好的纠错机制,组织者会经常在PG的网站上、Wiki界面和实时通讯界面发布消息,鼓励用户提供反馈并帮助校正、更新内容。同时,用户还可以通过RSS订阅新书服务,可以关注Twitter并通过Facebook和Google+分享内容。

4 技术层面

4.1 应用设备

PG是一个开放式文本文档数据库。它提供EPUB、Kindle、HTML和其他文本格式书籍(Project Gutenberg, 2011),支持PC、iPad、Kindle、Sony Reader、iPhone、Android和其他多种设备。PG采用ASCII简单文本或其他可修改的开放格式编辑文本,并采用不同的格式发布资源(PDF、HTML、XML等),只要电脑中安装有JavaScript及相关组件,用户就不必安装任何软件来阅读文献。同时,从PG下载的MP3格式内容也不需要特殊软件支持。

4.2 文件格式

为了使尽可能多的用户及其电脑设备流畅利用其免费电子书,PG使用简易通用数字化技术。PG的文献都被整理为ASCII格式,因为该格式可以被几乎所有的软硬件识别(Project Gutenberg, 2012)。图书馆的总索引文档的描述说明使用“最普通的格式”:文本文档、8-bit text、Big-5、Unicode、HTML、TeX、XML、MP3、RTF、PDF、LIT、MS Word Doc和PDB(Project Gutenberg, 2012b)。PG也将EPUB、Kindle纳入可用格式。比如,Jane Austen的*Emma*一书,有HTML、EPUB、Kindle(有影像)、Kindle(无影像)、PDF、Plucker、QIOO Mobile、Plain Text和UTF8等多种格式。

4.3 数据库结构和网络标记语言

据Newby表示,PG“输出XML和RDF,在后台使用Postgre SQL数据库及不同的自定义脚本(Perl、PHP和python)获取新标题”。其中,Perl是一种适用

于统一编码的开源编程语言 (Perl.org, 2012); PHP 即hypertext preprocessor (超文本预处理器), 是广泛使用的通用描述语言, 适用于网络发展且能被嵌入HTML语言 (PHP Group, 2012b); Python也是开源编程语言 (Python Software Foundation, 2011)。PG使用W3C的标记语言验证服务来核查文档, 每一个文档都设有链接指向其验证器, 其验证器则证实该网站UTF-8编码并融合XHTML和RDFa语言, 为浏览者呈现更清晰的XHTML编码内容。

4.4 其他技术服务

为了更进一步缩减数字鸿沟, 对于没有电脑或新式设备的人, PG提供可将所需内容复制到CD或DVD上为用户的定制服务。PG可为任何有信息需求的用户免费制作、赠送CD/DVD, 并鼓励用户为朋友、图书馆或学校制作备份。2003年8月, PG发布的最佳资料CD囊括了600本电子书, 当年12月为庆祝PG的第一个10000部电子书发布了第一个DVD, 并烫印了9400部书名。

5 不足之处及有关建议

PG的优势显而易见: 用户易用性好, 无附加费用, 版权信息完善, 下载数据清晰, 用户反馈及时快捷, 资源更新较快, 权威性高, 收藏价值较高等。用户首次使用时即可快速定位网站并选定自己想要的文件格式。它从全世界范围征集尽可能多的资源和帮助, 为尽量多的读者提供不同语言不同类型的免费信息资源, 对数字图

书馆的发展影响重大。但其在提供信息服务的过程中也不可避免地存在一定问题。

首先, PG的缺陷在于其视觉效果不佳。用户交互界面较陈旧, 网页设计缺乏网络标记语言, 缺乏吸引力。网站检索结果较差, 缺乏“联想”功能。界面层次不清晰, 子目录和书架(Bookshelf)内容交叉, 用户必须多次点击鼠标才能进入记录。

其次, PG无明确目标或管理机制引导馆藏发展, 馆藏的增加多依靠志愿者个人意愿。因此, 图书馆的收藏文献失衡, 如小说的数量远远超过非小说内容, 原因是大多数志愿者愿意完成小说类文献的数字化。同时, 确定的员工和有效的计划管理对于数字图书馆来说举足轻重, 而PG的发展主要依靠志愿者的推动而不是由PG主动设定目标并选择志愿者完成工作, 这可能影响PG的稳定和持续发展。

最后, 版权限制是PG发展的最大障碍。PG严格遵守版权法, 任何一部书籍在进入获取领域之前都要经过版权审核, 这意味着等待资源进入公共领域需要很长一段时间。因此除了经典读物, PG提供的大部分内容可能并非普通读者喜欢的内容, 内容信息性强但存在过时现象。

总而言之, Michael Hart启动了第一个全球性数字图书馆。PG为全世界提供长期免费的开放资源, 对于数字图书馆的资源建设具有较强的影响力。然而, PG在将来的发展过程中, 需要更精美时尚的网页设计, 提高PG对用户的友好程度和服务娱乐性, 收集更丰富多彩的数字资源, 提高PG的社会影响力和信息服务能力。

参考文献

- [1] Distributed Proofreaders. DP: Welcome [EB/OL]. [2012-01-31]. <http://www.pgdp.net/>.
- [2] WELLER S. Project Gutenberg sets 10,000 book goal [J/OL]. Publisher's Weekly 24/4 (2000-04-03). Infotrac, 2000.
- [3] iBiblio.org. About [EB/OL]. [2012-02-05]. <http://www.ibiblio.org/about/>.
- [4] Librivox.org. Complete FAQ [EB/OL]. [2012-01-31]. <https://forum.librivox.org/viewtopic.php?f=18&t=219&sid=7fd6856af89f3c6bf2905c4d52fb0ce8>.
- [5] PHP Group. PHP: General Information - Manual [EB/OL]. [2012-02-05]. <http://us3.php.net/manual/en/faq.general.php>.
- [6] Project Gutenberg. Project Gutenberg needs your donation [EB/OL]. [2012-01-31]. http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Needs_Your_Donation.

- [7] Project Gutenberg. Project Gutenberg online EpubMaker [EB/OL]. [2012-01-31]. <http://epubmaker.pglafl.org/>.
- [8] Project Gutenberg Literary Archive Foundation. Cease and Desist Responses [EB/OL]. [2012-01-31]. <http://cand.pglafl.org/>.
- [9] Python Software Foundation. About python [EB/OL]. [2012-02-05]. <http://python.org/about/>.
- [10] VAKNIN S. The Ubiquitous Project Gutenberg interview with Michael Hart, its founder [OL]. [2012-02-01]. <http://samvak.tripod.com/busiweb46.html>.
- [11] Project Gutenberg. About: Gutenberg [EB/OL]. (2008) [2011-02-03]. <http://www.gutenberg.org/wiki/Gutenberg>About>.
- [12] HART M. The History and Philosophy of Project Gutenberg. Project Gutenberg [EB/OL]. (1992) [2012-01-31]. http://www.gutenberg.org/wiki/Gutenberg:The_History_and_.
- [13] HART M. Project Gutenberg Mission Statement [EB/OL]. (2007) [2012-01-31]. http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_.
- [14] NEWBY G. Anacleto Search of Project Gutenberg's eBooks. Project Gutenberg News [EB/OL]. (2007) [2012-01-31]. <http://www.gutenbergnews.org/20070820/anacleto-search-of-project-gutenbergs-ebooks/>.

作者简介

刘燕权，男，博士，毕业于美国麦迪逊大学信息图书馆学院，现在南康涅狄格州立大学任教。他近年的出版物主要集中在数字图书馆、数字化及多媒体技术、软件管理、国家信息基础结构、信息存储、图书馆统计及管理。E-mail: liuscsu@gmail.com

刘晓东，女，硕士研究生，现就读于北京师范大学管理学院图书馆学专业。E-mail: bnugyliuxiaodong@126.com

Project Gutenberg – The Forerunner of Digital Libraries

Yan Quan Liu / Southern Connecticut State University, New Haven, CT, USA, 06515

Liu Xiaodong / Beijing Normal University, Department of Management, Beijing, 100875

Abstract: Project Gutenberg has now been universally recognized as the very first informational provider on the internet and the oldest digital library. Project Gutenberg started on July 4, 1971, Michael Hart typed the content of the U.S. *Declaration of Independence* into a computer, and began a collection of electronic books that built up the Project Gutenberg. With countless volunteers' contribution, the project currently boasts over 42,000 free eBooks on its site and over 100,000 eBooks through its affiliated organizations after 40 years. Hailed as the first digital library, Project Gutenberg has played an important role in the construction and development of digital libraries. This paper explored the project based on its background, collection organization, services and technologies features. Authors' comments were also given.

Keywords: Digital library, Project Gutenberg, Free E-book collection, Digitization collection

(收稿日期: 2013-03-05)