大型数字化项目的范围限定与术语辨析※

宋琳琳[1] 黄如花[2]

([1]武汉大学信息管理学院 武汉 430072)

([2]武汉大学信息资源研究中心 武汉 40072)

摘要： 本文首先划分大型数字化、大规模数字化、非大型数字化等数字化项目的类型，并列举各类型的相关术语。通过对术语的辨析，得出MD界定的关键为是否存在资源选择，和参与MD的目的。然后从MD的开展情况入手，调查各MD项目的开展动机和资源选择标准。文章认为，资源选择标准是大型数字化项目必要组成部分。

关键词： 大型数字化项目 资源选择标准 大范围数字化

中图分类号： G250.76

图书馆从20世纪90年代中期开始从事大量馆藏纸质文献的数字转化工作，期间著名的项目有古登堡计划、百万图书项目（Million Book Project, MBP）等。2004年，由Google与哈佛大学图书馆等5个图书馆(G5)合作开展的Google 图书搜索（Google Book Search, GBS），使得大型数字化项目飞速发展。2005年建立的开放内容联盟（Open Content Alliance, OCA）也拥有众多的合作图书馆，如多伦多大学图书馆、加州大学图书馆、波士顿图书馆联盟等，同时由惠普实验室、微软、雅虎、Adobe等公司提供技术和设备支持；微软还于2006年研发了专门用于图书检索的搜索引擎Live Book Search。为回应GBS，欧洲于2006年开始建立欧洲数字图书馆（European Digital Library, EDL），并于2008年10月投入使用，EDL对来自欧洲各国的文化遗产进行大范围数字化，通过网络提供多语言的一站式服务。

1 数字化项目的类型划分

尽管数字化项目已广泛开展，但其类型划分却始终没有统一标准，尤其一些相关术语，如大型数字化（Mass Digitization, MD）、大规模数字化（Large-Scale Digitization, LSD）等有待进一步明确区分。

1.1 大型数字化

Karen Coyle提出，MD是比LSD范围更大的一种数字化项目。它以工业化生产模式对资料进行数字转换，是将整个图书馆馆藏不加筛选地转换。MD的目标不是创造馆藏而是全部数字化，即数字化所有印刷型图书。为了更经济有效地实现这个目标，MD需要高效率地扫描每页图书，通过OCR识别这些扫描图片形成可检索文档。人工干预被降到最低，这样OCR生产出的可检文档就可作为检索资源使用而不必再进行修改。当然，MD也需要一些人工操作，如添加页码，表格内容等，因为这些无法通过OCR自动生成[1]。

Karen Coyle对MD的界定，在业内广受支持，加州大学数字图书馆（California Digital Library, CDL）的大型数字化项目在网站的FAQ中直接引用了该定义。“密歇根大型数字化项目影响的高峰论坛”中，Clifford Lynch认可通过范围划分数字化项目，称“图书馆、档案馆和博物馆对其馆藏的古老文献基于保存和检索目的进行的数字化称之为大规模数字化更为合适，而不是大型数字化；因为现在和未来会进行不同规模的数字化项目，而上述机构的规模还达不到‘大型’的要求”[2]。

1.2 大规模数字化

LSD与MD相比，最大的特色在于LSD是选择性数字化。LSD同样生产大量的扫描页面，但其关注馆藏建立并生成一系列文档，用于数字文本的检索和保存。而其他方面，如扫描图书的数量、扫描速度等方面，LSD与MD并无本质区别。LSD项目的典型案例有美国国会图书馆的“美国记忆”，美国国家科学基金赞助的“百万图书项目”，还有很多大型图书馆开展的数字化项目等。

1.3 非大型数字化

与 MD 相反的是那些对数字化资源进行精心挑选的数字化项目，称为非大型数字化。

※本文为教育部人文社会科学重点研究基地重大项目“数字信息时代的图书馆管理研究”（07JJD870221）成果之一。

这是基于保存目的兴起的数字化项目，为日益恶化的文献生产替代品，或为使珍贵的资源得到更广泛应用。如维吉尼亚大学图书馆的 E-TEXT project 和 Adobe 公司的 Octavo Editions。非大型数字化的另一个特点是它的最终产品。MD 的最终产品是对图书页面的扫描，并以可检索的 OCR 形式进行备份，缺少深层的信息加工组织；而非大型数字化的终端产品则生成大量已标识的文本，提供大规模的使用。

1.4 MD 与 LSD 合为一类

此类划分，认为 MD 与 LSD 之间没有本质区别，承认资源选择标准存在的必要性，建议模糊两者的区分标准。Clifford Lynch 认为对于图书馆的藏书来讲，没有全部和最终的概念，图书馆的馆藏包括数据库、手稿、音乐、图片、多媒体等各种形式，采用 LSD 的概念反而更容易清晰分辨用户的需求，建议用 LSD 代替 MD，因为相较于 MD，LSD 更适合规划[3]。Conway 认为，MD 就是一系列 LSD 项目的累积，是为社会大众进行的数字化[4]；CDL 的 John A. Kunze 提出“所谓大型数字化，就是在世界主要的图书馆对报纸、图书、视频等文献类型进行大规模的扫描”[5]；而美国图书馆与信息资源委员会(CLIR)将 GBS、OCA、MBP 等项目划归为 LSD 项目，直接将 MD 与 LSD 融为一体[6]。

2 MD 的特征辨析

2.1 MD 的公认特征

MD 具备两个公认的显著特征，即海量图书及数字化和元数据创建中的高速扫描。

快速、高效和高度自动化经常被用来形容 MD 项目的扫描和元数据创建，这些项目均包含大量图书，如 Google 曾计划与 G5 合作实现 1500 万册的图书扫描计划；与美国机构协作委员会（Committee on Institutional Cooperation, CIC）合作数字化 1000 万册图书。而 EDL 的目标是在未来几年至少完成 600 万册图书的数字化。

MD 的扫描速度也是惊人的，OCA 在 3 个国家拥有 8 个数字化中心，每个数字化中心都拥有 10 台高速扫描仪，每天工作 16 小时，大约每月可以扫描 12000 册图书[7]。密歇根大学图书馆曾预测若与 Google 合作，该馆的 700 万册图书将在六年的时间内完成数字化，而未加入 GBS，按其原先速度完成相同数量大约需要 1000 年[8]。

2.2 MD 界定的分歧

通过对上述类型划分的比对，可见 MD 界定的关键在于：一，是否存在资源选择标准。Google 誓将全球图书数字化，同时也有学者认为“MD 不应进行资源选择，因为筛选会增加数字化成本，放慢工业化进程，这就违背 MD 项目的初衷。”二，MD 项目参与者的动机。商业公司、图书馆、非营利组织参与 MD 的动机各有不同，这主要由自身性质决定，其虽不能作为类型划分的标准，但可作为区分参与者性质的参考。

3 MD 项目参与者的动机分析

表 1 MD 项目主要参与者的动机[9]9

参与者	首要目标	区分处
研究图书馆	保证丰富知识的机构核心目标实现 改变用户获取图书馆资源的方式 确保图书馆资源在未来人家可以检索 利用数字版本备份 发展先进检索工具、进校文本挖掘实验	保持其在信息收集、管理、保存、检索方面的地位，以促进学习、教学和研究
GBS	对多语言公共领域的图书提供完整可检索的索引 方便公众利用 Google 进行图书检索和发现 通过提供优秀的搜索引擎吸引更多的用户	对全世球的图书数字化以便深度检索的开展
Microsoft	建立全文图书馆数据库 方便公众轻松找到相关图书 采用特有的界面传送结果并为检索提供先进工具	通过对授权内容建立可信赖的索引实现网络检索向信息检索的转变
OCA	建立开放存取的数字馆藏 支持多语言、多媒体的数字文本的永久保存 通过多种存储设备保存文档	建立学术资源的长期保存，并对所有搜索引擎开放
MBP	提供用户利用网络对高质量资源快速、方便的检索 确保全球的用户都能合理的利用数字化资源 为信息存储与管理、搜索引擎、机器翻译等提供实验基地	为 MD 的检索和管理提供试验基地

根据 Rieger 的调查，所有参与者的主要动机都是为了方便图书的检索。但由于不同参与者的性质不同，参与目的也呈现多样化。以图书馆为例，其目的包括检索、保存、支持科研 3 方面，但任何一类目标的实现都面临困难。数字化方便了用户对图书馆资源的检索，但受资金、图书版权等限制，网络的公开检索成本太高，难以实现。生成数字馆藏或作为替代品也是图书馆的目的之一，但是由于技术等限制，数字化产品质量不高，很难完全替代原有馆藏；随着技术发展，现有的数字馆藏在未来能否继续使用也不能保证。因此，参与动机难分伯仲，不能作为类型区分的标准，只能作为区分参与者性质的参考。

4 MD 项目的资源选择情况

4.1 提交数字化的馆藏数量有限

GBS 项目开始之初，曾宣言要对全世界的图书进行数字化，但从项目开展的情况看却难以实现。G5 成员之一的哈佛大学图书馆拥有 1580 万册图书，但只提供 100 万册图书供 Google 扫描。其他合作馆提供 Google 数字化的图书基本上少于馆藏总量的 1/10，这种馆藏图书的选择性提供必然需要严格的资源选择标准进行筛选。

表 2 部分 GBS 合作图书馆资源选择情况

合作馆	牛津大学	哈佛大学	维吉尼亚大学	哥伦比亚大学	康奈尔大学	CIC
合作时间	2005	2005	2006	2007	2007	2007
馆藏数量/拟数字化数量	1100 万 /100 万	1580 万 /100 万	720 万/50 万	920 万/100 万	800 万/50 万	78898 万 /1000 万
版权限制	公共领域	无	公共领域	公共领域	无	无
特藏优先	否	否	是	是	是	是
存在选择标准	否	否	是	是	是	是

4.2 拥有资源选择标准

牛津大学图书馆与 Google 合作伊始，双方协定尽可能多得对馆藏数字化，缺乏资源选择机制。既便如此，该馆也意识到并非所有馆藏都适合数字化，规定脆弱文献如手稿、档案、地图、早期印本图书应排除在数字化范围之外[10]。随着项目的深入开展，后续与 Google 签约的图书馆都制定较明确资源选择标准。与 GBS 相比，OCA、MBP 和 EDL 都制订较详细的资源选择标准。OCA 将选择标准的制定权交给各合作馆，只对提交的数字化资源进行有用性测试和筛选。EDL 则制定了适应其需求的不同类型资源的选择标准。

总体来讲，已制定的资源选择标准涉及学科领域、图书形态、版权、质量控制、特色馆藏等方面。首先是对特色馆藏和高利用率资源优先数字化；威斯康辛大学-麦迪逊分校图书馆主要将该馆高利用率的文献如医学历史、发明专利、工程历史、地方文献、地图和活页乐谱等内容进行数字化[11]。波士顿图书馆联盟的成员馆主要是数字化其特色馆藏，如布兰代斯大学图书馆的德雷福斯事件的文献，波士顿大学图书馆的非洲方面资料，波士顿公共图书馆的家谱资料等[12]。其次，图书的学科范围也有要求；康奈尔大学规定其 Mann 图书馆涉及生物科学、环境科学、应用经济、公共政策及管理、纺织科学、营养与食品科学的馆藏将被优先数字化[13]。微软要求耶鲁大学图书馆向其提供的图书应限定在艺术、艺术历史、历史、宗教学科领域 [14]。版权方面，OCA、EDL、MBP 三个项目均以数字化公共领域资源为主，若要对受版权保护的资源数字化，则必须征得版权持有者的许可。牛津大学图书馆不仅规定只对公共领域不受版权保护的图书数字化，同时还将图书出版年限提前到 1885 年。图书形态上，耶鲁大学图书馆规定图书单独页面的宽度小于 9.7"，高度小于 14.5"；图书至少能张开 75°；必须是侧面或上面装订，装订线附近必须有 1/4"的边距[15]。

4.3 资源选择标准多样化

MD 项目由来自全球的文化技术组织、非盈利机构、商业公司和政府部门合作开展，为平衡各方利益，资源选择标准必然多样化。以 MBP 为例，其合作方涉及美、中、印、埃、卡塔尔等国，面对各国合作馆的大量馆藏图书，选择标准因国不同存在差别。美国方面，主要是各成员馆自行制定标准选择联邦政府和各州政府的政府出版物，以及各类科技报告和论文；印度方面主要数字化印度成员馆选择的 11 种官方语言著成的政府出版物。在中国，该合作项目被称为“高等学校中英文图书数字化国际合作计划”，其资源选择标准区分为中、英文两种类型。拟数字化的英文图书由美国数字图书馆联盟协商决定。中文资料则采用学位论文全部数字化；现代图书采取按出版社等级分配，各单位自行选择分配的拟数字化的图书，

并提交专家委员会审核；传统文化资源由各参加单位选择特有的古籍和传统文化资源进行数字化，并将拟数字化的书目或内容上报项目管理中心，经批准后进行数字化，形成专题数据库等形式[16]。

4.4 图书馆在资源选择中发挥主导作用

作为资源的收藏和提供者，图书馆了解各种资源的价值，也掌握用户需求，是资源选择的最佳执行者。OCA 提出“参与馆根据自己的馆藏情况自行决定拟数字化的图书”；PALINET 作为机构联盟加入 OCA，规定其数字化的图书由各个图书馆自行挑选，只要这些图书不受版权保护即可，同时拟数字化的图书还需拥有基本的元数据；但其强烈建议选择特殊的、地区性的文献以建立地区性的历史收藏，同时也降低重复建设的可能[17]。EDL 对于资源的选择并没有推行“自上而下”模式，鼓励各参与馆自行制定资源选择标准，但应符合 EDL 对不同类型文献的基本要求。EDL 要求参与馆提供的图书必须属于其馆藏，还应该符合人力和财力平衡协调的原则。所以，各参与馆大多提供具有本国特色、本民族文化价值的资源[18]。

4.5 成立专门资源选择机构

MD 项目的参与馆十分重视其在资源选择中的关键作用，除制定严格的选择政策外，很多图书馆还建立专门的资源选择机构，调动全馆、全校师生参与到资源选择中来。CDL 为了实现与 OCA 和 Google 的合作，专门成立了大型数字化馆藏选择顾问委员会。其任务包括为 CDL 建立一套通过数字化扫描进行馆藏回溯和选择的内部机制，为潜在的扫描资源制定评价标准，同时负责与 CDL 工作人员，UC 编目员进行沟通，向该馆的项目和技术部门寻求数字化过程中的技术和编程的帮助。为馆藏发展部门提出 MD 项目在馆藏发展中应注意的问题，并向其推荐图书等。新汉夏普大学图书馆鼓励本校师生员工推荐希望扫描的图书，被推荐的图书如若符合该馆的 OCA 扫描标准，就会被优先扫描[19]。

5 结论

通过上述对 MD 项目的调查及分析，笔者认为：首先，资源选择是 MD 项目的必要组成部分。由于版权、图书形态、重复建设等限制和各参与馆资源建设的需要，数字化所有图书难以实现；正在开展的 MD 项目也印证了资源选择标准的可行性和必要性。所以，MD 项目必须制定资源选择标准，具体包括学科范围、优先数字化机制、图书形态描述、图书推荐方式等各方面。其次，由于 MD 项目的参与主体不同，其性质存在差别，这就导致各项目的实施动机各异。所以图书馆在合作开展 MD 项目时，应辨明合作方的动机与最终产品，以便合理规划。

参考文献

[1] Karen Coyle. Mass Digitization of Books. The journal of Academic Librarianship, vol.32, no. 6, November, 2006, pages 641-645.

[2]NCLIS. Mass Digitization: Implications for Information Policy [C]//Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects, 2006. University of Michigan

[3] Clifford Lynch. What is mass digitization. [EB/OL]. [2008-12-23]. http://infomotions.com/musings/mass-digitization

[4] Paul Conway. Tec(h)tonics: Reimagining preservation. C&RL News, November 2008, Vol. 69, No. 10

[5] John A. Kunze. Where Preservation Meets Mass Digitization. [EB/OL]. [2008-12-6]. http://lauc.ucmercedlibrary.info/lauc_mass_dig.ppt

[6], [9] Oya Y. Rieger. Preservation in the Age of Large-Scale Digitization .CLIR 1755 Massachusetts Avenue, NW, Suite 500 Washington, DC 20036

[7] OCA [EB/OL]. [2008-12-12]. www.opencontentalliance.org/

[8] Michigan Digitization Project [EB/OL]. [2008-12-18]. http://www.umich.edu/news/index.html?BG/google/index.

[10] Oxford-Google Digitization Program [EB/OL]. [2008-12-20]. http://www.bodley.ox.ac.uk/google/

[11] University of Wisconsin - Madison Google Digitization Project [EB/OL]. [2008-12-25]. http://www.library.wisc.edu/digitization/#about

[12] BLC-OCA project [EB/OL]. [2008-12-15]. http://bc.edu/libraries/newsletter/2008summer/jesuit/index.html

［13］ Cornell University Library-Google Library Partnership［EB/OL］. ［2008-12-28］. http://www.library.cornell.edu/communications/Google/faq.html
［14］，［15］ Jennifer Weintraub and Melissa Wisner .Mass digitization at yale university library—exposing the treasures in our stacks［EB/OL］. ［2008-12-28］. http://www.infotoday.com
［16］ CADAL 资源选择标准［EB/OL］. ［2008-12-25］. http://www.cadal.cn/
［17］ PALINET-OCA project［EB/OL］. ［2008-12-27］. http://www.palinet.org/dsfaq.aspx
［18］ EDL［EB/OL］. ［2008-12-28］. http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/05/347
［19］ Open Content Alliance Submission Criteria［EB/OL］. ［2008-12-29］. http://www.library.unh.edu/diglib/criteria.shtml

The Analysis of Breadth and Glossary in Mass Digitization

Song Linlin[1] Huang Ruhua[2]

([1]School of Information Management, Wuhan University, Wuhan 430072)

([2]Center for the Studies of Information Resources，Wuhan University, Wuhan 430072)

Abstract： This paper introduces different types of digital projects such as Mass digitization, Large-scale digitization, and lists some related glossaries. Then it compares these glossaries and finds that the difference among them is whether the resource selection criteria exist and their motivations. It also surveys some MD projects and their resource selection criteria, such as Google, OCA, and MBP to find that resource selection criteria are the necessary part in mass digitization.

Keywords: Mass digitization, Resource selection criteria, Large-scale digitization