●臧国全

论图书馆信息资源数字化项目成本节约*

摘 要 图书馆信息资源数字化需要资金的高投入,应当节约成本。其成本主要包括:数字化内容选择,数字化生产准备,元数据析出,原始文献保护,原始资源替代品生产,基础设施构建,数字转换,文本抓取等。节约成本的主要方法有:减少人力成本,项目外包,自动化,优化生产流程,强化质量管理。表 2。参考文献 8。

关键词 信息资源 数字化 项目成本 成本节约
分类号 G250.76

ABSTRACT Library information resource digitization projects need high investment and also require cost saving. The major costs include digitization content selection, digitization production preparation, metadata analysis, source document preservation, surrogate production, infrastructure construction, digital conversion, text capture, etc. Major measures for cost saving include human resource reduction, project outsourcing, automation, production work flow optimization and quality control. 2 tabs. 8 refs.

KEY WORDS Information resource. Digitization. Project cost. Cost saving.
CLASS NUMBER G250.76

1 项目成本构成

图书馆信息资源数字化是一类比较复杂的项目工程,生命周期包括项目规划、数字化内容选择、生产准备、数字化生产、元数据析出、数字保存和数字资源发布等主要阶段。每个阶段的成本构成不同,在项目规划阶段对整个项目进行成本预算时,需要考虑的要素有很多。已经有一些案例研究了数字化项目成本的测算,并对成本构成要素进行了探索,这些大都被编制在美国国家网络文化遗产项目(NINCH)的资源列表中[1]。

美国研究图书馆集团(RLG)设计了一个比较实用的数字化生产成本构成要素工作单,以帮助图书馆制订数字化项目的经费预算。这个工作单包括了 10 个步骤和成本要素[2]:数字化内容选择、进行数字转换的信息资源集合规模和类型、原始信息资源数字化前的准备、数字扫描标准的建立、元数据方案的确定和元数据内容的生成、数字扫描成本估算、数字化文本转换成本估算、采用标记语言对数字文档进行编码的成本估算、数字文档的后期处理成本和其他成本估算。

欧盟文化与科学信息资源数字化部长级网络(Minerva)在其 2006 年 6 月发布的研究报告中,依据生命周期,将信息资源数字化整个项目的实施成本分解为下述 9 个主要因素[3]。

(1)数字化内容选择。信息资源数字化内容选择是依据确定的标准进行相符性判断,将符合条件的原始资源遴选出来,继而进行数字化加工的过程。这个阶段的成本体现在内容选择标准的构建和依据选择标准对馆藏文献进行筛选两个方面。

(2)数字化生产准备。Minerva 通过调查估算,该阶段的成本占项目总成本的 20% ~30%。成本的主要构成有:

原始文献载体搬运,涉及到对搬运文献的编目和打包。

原始文献唯一标识符的添加。该标识符有助于资源编目,并可为以后扫描和元数据抓取提供便捷。

原始文献的状态评估,以确定合适的搬运、处置和数字化加工的方式。

原始文献的数字化加工准备,包括去除原始资源载体的装订,以及对载体表面的清洁处理等。

版权状态和其他使用权的明晰以及版权许可,以免日后图书馆被起诉,导致信誉损失甚至被迫取消电子版。

残缺文献的处理。对残缺不全的馆藏文献在数字化之前要进行补缺或做相应说明。

原始文献返回的检查,确保其完整和未被损坏。

(3)元数据析出。数字化项目的元数据不仅要

* 本文为国家社科基金项目"图书馆信息资源数字化建设模式研究"(编号 05BTQ007)研究成果之一。

对数字对象的内容进行描述,还要对其加工过程、采用的技术和工艺、产权管理等事项进行描述。析出方法以人工为主,软件自动抽取为辅。

(4)原始文献的保护。对于数字化过程可能造成载体损伤的文献资源(如易碎载体),在数字加工过程中要采用合适的技术和措施进行保护。

(5)原始资源替代品(如胶片)的生产。原始资源替代品的作用有二:其一是用于长期保存;其二是替代原始资源进行数字扫描,这种扫描方式的成本要比直接对原始资源进行扫描低得多。一般来说,信息资源数字化项目的产品包括原始资源的替代品和数字产品,前者用于长期保存,后者用于数字存取。但这种替代品的生产是需要成本的。

(6)基础设施的构建。包括网络、数据存储、备份、数字资源发布、软硬件等。

(7)数字转换。包括数字扫描、数字拍照、模拟音频和视频信息资源的数字转换。

(8)文本抓取。采用 OCR 或重新键入等方式对文本图像进行识别,也包括嵌入标记语言(如 XML)的标识。

(9)整个项目实施的质量控制。

2 项目成本节约

数字化项目成本节约的"黄金法则"是审查需要人工干预的各个环节,并尽可能取消或减少这些环节。降低数字化项目成本的基本思路是:减少人力成本;对数字转换和元数据析出等环节实施自动控制,以减少人工干预;实施规模生产,减少生产流程中的变量;提高整体绩效和产出,确保资金高效利用;严格质量管理,改善和优化项目生产流程;构建风险管理预警体系,降低风险成本;培训员工所需技能,提高生产能力和产品质量;开展项目合作,实现资金、设备、人员和技术的优势互补。下面是节约成本的主要方法。

2.1 减少人力成本

信息资源数字化项目的人力成本不仅包括基本工资,还包括其他成本因素。Minerva 采用的快速估算人力总成本的方法是基本工资乘以 170%。降低工资成本通常意味着寻找较廉价的劳动力,雇用较低技能的员工。实现的方法有:

(1)把任务分解成若干模块。对于复杂的数字化项目,根据所需技能知识的类型与程度,将工程任务进行详细分解,分块设计,使较低技能的员工从事重复工作,当绝对必要时才聘用高技术专家。

一个例子是欧盟的一个植物样本数字化项目。该项目不仅涉及数字化技术知识,还涉及植物学知识。植物样本图像扫描可以从整个工程中分解出来,作为一项重复性工作可由经过培训的工资较低的人员来完成;数据库录入工作需要一定的植物学知识,但可以通过设计一套规程(如从植物样本的注解中提取元数据内容的规范)由简单脑力劳动者实现,仅需雇用为数不多的植物学专家对较难的情况提供建议和指导。

(2)提供良好工具和指导。一般认为,数字化项目需要专业技术人员通过操作高技术性能设备来完成,是高人力成本项目。实际上,除了把任务分解为复杂程度不等的模块,良好的工具、恰当的指导和操作手册,同样可降低对诸多任务的技术要求,从而降低工资成本。比如在数字图像转换时,颜色管理软件的使用可以准确校对数字化设备和环境,保证成像颜色的准确度,这样可以减少对相关技术专家的依靠性。另外,在元数据建立一套数据抓取范围表也可以降低对专业知识的要求。使用软件工具对录入的元数据自动添加 XML 标记,同样可以降低任务的复杂度,并且还可提高最终产品的质量和标准化程度。

(3)培训投资。培训投资似乎有悖于通过降低技术需求减少成本的原则。但事实上,对较低技能的员工进行适当的针对性培训能够使他们在一定范围的工作中具备与高技术人员同样的表现。

根据 Minerva 的统计,数字化项目中有 85% 的任务属于重复性工作,12% ~15% 的任务比较困难,只有小于 1% 的非常困难。一般情况下,雇用技能较低的员工,对他们进行培训,可以胜任 12% ~15% 那部分工作,剩余任务可通过独自承担或项目外包的形式由专家解决。培训投资还可以改善工作流程,提高生产量,并降低由于质量问题而重做的成本。

2.2 项目外包

寻求较廉价的劳动力是数字化项目进行外部代理的重要原因。对于大规模数字化工程,外包比图书馆自己承担要节省成本。外包对于不具备数字化生产基础设施的图书馆也具有很强的吸引力,因为这样可以省去设备购置的高昂成本。对于一些特殊载体文献(如大幅面建筑图纸、地图和海报等)进行数字化,外包也许是唯一选择。对于短期数字化项目,外包尤其是比较理想的选择。然而对于长期数字化项目,完全外包值得商榷,将整个项目的实施依赖于外包商并不明智。外包方式是图书馆的困难选择,决策

前应充分咨询,出发点是成本效益分析。

表1是外包和自己实施数字化项目的优缺点对比分析,图书馆在进行实施方式的选择时,应该全面权衡利弊。

表1 外包和自己实施的对比

	图书馆自己实施	外包
优点	边学边做,逐步培养自己的数字化专家。 发展自己的生产能力。 可对数字化生产过程的所有环节进行控制。 定义需求具有灵活性。 原始信息资源的安全可以得到保障。	外包商拥有专家和训练有素的员工。 可以根据项目规划和预算,基于被数字化的原始文献的规模,谈判每件文献数字化的价格。 劳动力成本低。 技术过失成本由外包商吸收。 可供选择余地大。
缺点	投资大。 每个图像的生产成本不具有可议性。 需要构建基础框架,包括生产场地、数字化设备和计算机等。 生产能力和效率有限。 容易产生技术过失成本。 对图书馆的其他活动会产生影响。 除了生产费用,还需要支付设备、维护和人力费用。 需要专门技能的员工,并对员工进行培训。 需要设备支持。	图书馆在整个数字化项目实施进程中,缺少了一个重要步骤。 外包商可能对图书馆的需求不熟悉。 图书馆不能进行现场质量控制。 外包商生产的数字图像可能还需要图书馆来加处理,图书馆应该对生产的图像随机抽取检查。 需求必须在合同中清晰定义,否则有可能出现交流障碍。 原始信息资源需要搬运,并由此导致运输安全和处理安全问题。尤其对于三维型的原始资源载体。 容易受外包商的稳定性影响。

外包不仅意味着使用第三方提供的设备和专家,而且意味着不承担维护专业设备的成本。实际上,很多数字化项目都没有充分使用所购买设备的所有折旧价值,变相增加了数字化生产的成本。

从整体上考虑,为了权衡外包的合理性,以下问题可供图书馆权衡:通过外包,图书馆能否专注自身核心技术的提升,并利用外包商的专长来降低项目的总成本;外包能否提高图书馆的竞争力,能否强化项目目标的实现;能否增加图书馆与关键技术和设备的接触机会,而这些技术和设备都是图书馆购买不起的;能否降低技术过时的风险成本;从人力资源角度,能否提高规模经济;能否获得和增加与技术人才相接触的机会;外包能否强化对项目费用的控制。

2.3 自动化

数字化项目实施过程中自动化程度越高,成本降低的幅度就越大。针对这类项目,"自动化"包括两种类型:其一是机械自动化,减少乃至完全取代原始资源处理过程中的人工干预;其二是基于软件的自动控制,提高生产效率,取消人工干预,提供与员工的交互界面过程,反而会导致一些瓶颈。比如目前大多数的数字化项目中质量检查和元数据析出两个工序大都由人工完成,其他工序的自动化就很可能导致这两个工序成为整个生产流程的瓶颈。

2.3.1 机械自动化

机械自动化主要应用在数字化项目中数字扫描阶段。影响该阶段成本主要有两类任务:扫描过程中处理和移动原始资料的成本以及对输出图像文件标注必要说明的成本。这两项工作占用时间越多,成本越高。除了生成特大文件(大于300Mb),技术成本微不足道。因为从整个生命周期角度看,通过改善计算机硬件(使用可移动存储器、存储局域网络和光纤网络等)而产生的存储技术成本在整个项目的实施成本中所占的比重很小。

自动化可以降低数字扫描成本,降低程度与自动化水平密切相关。例如:使用纸张供应器可使单页文献自动扫描;智能扫描仪可扫描装订成册的纸质文献;幻灯片可装载在拖盘里进行批量自动扫描;缩微胶卷虽可以自动扫描,但单片缩微胶卷和有包装的胶卷需要人工干预。

自动扫描是近年的新发展,可大大降低装订卷册的扫描成本。虽然不适用于中世纪手稿,但除了在扫描开始将其放置在机器上,19世纪的装订本可以不加人为干涉地被扫描。目前投入使用的自动扫描设备主要有:4DigitalBook数字化流水生产线[4],声称可以每小时扫描1500到3000页;Kirtas自动书籍扫描系列[5],每小时扫描2400页;Atiz BookDrive桌面自动翻页扫描仪[6],每小时也可扫描500页。还有一些半自动化的针对装订卷册的扫描仪,如Zeutschel和12S等。

自动化实现的低成本是建立在高成本数字化设

备的基础之上。对于小规模项目,除非考虑外包,否则高端的扫描设备购买可能从成本效益角度不合适。

2.3.2 *基于软件的自动化*

基于软件的自动化目的在于加速生产处理过程。在数字化项目实施过程中,这类自动化的最好例证是图像文件的批处理(如图像编辑和基于数字化主文档的各种副本的生产)和文本内容的自动识别。批处理具有较高的成本效益比。实现对数字图像进行批处理的软件工具有很多,价格较低的有 Adobe Photoshop,比较昂贵的有 12S Bool Restorel 等。但到目前为止,软件自动化在下述 3 个领域对于成本节约的效果很有限:其一是基于多媒体的元数据析取;其二是文本资源的 OCR 识别,有时重新键入更经济;其三是智能元数据抓取,基于内容的描述性元数据抓取仍需要高水平的专业技术人员来实现,并时常需花费大量时间。

文本抓取是将依附于物理载体的信息内容转换为计算机可识别的形式。实现文本抓取的主要方法有手写识别、重新键入、语音识别和 OCR。

手写识别是一种先进技术,但 Entlich 对该技术的评论是[7]:对于手写体文本,目前还没有识别率能被用户广泛认可的商业软件,将其识别为可索引的数字文本。并建议,由于手写形式的局限性,在其识别技术实现根本性革新之前,谨慎用于规模自动化。

重新键入大都是一个人工过程,只有少部分可以借助软件提高效率。重新键入可以产生非常准确的结果,但人工成本相对昂贵。如果要求高准确率(比如 99.99%),OCR 可能比人工键盘输入成本更高,因为校对 OCR 文本要比借助廉价帮助软件进行人工键盘输入要昂贵。

语音识别是借助软件将音频转换为数字化文本的过程。目前提高语音识别准确度的常用方法之一是对语音输入者进行发音训练。然而,作为一项成本节约技术,基于音频内容的自动识别目前还不成熟。

OCR 是数字化文本生产中降低成本最明显的工具。它被用来自动从数字图像中识别生成文本,准确率与数字图像的质量有关。

选择成本最节省的文本抓取方案有时很困难。表 2 是一个文本抓取决策矩阵,总结了文本抓取的较合适方法[8]。

表 2 文本抓取决策矩阵

目的	被抓取的文本数量和类型	OCR 识别且无需校对	OCR 识别且需要校对	重新键入
全文抓取或用于自动生成索引	被抓取的文本图像数量小于 100			★★[注 1]
抓取目的:仅用于自动生成索引;文本印刷时代:Modern[注 2]	被抓取的文本数量和类型不限	★★	★	
抓取目的:仅用于自动生成索引;文本印刷时代:Historic[注 3]	被抓取的文本数量和类型不限	★★	★	
抓取目的:全文文本数字转换;或对手写文本自动生成索引	被抓取的文本数量和类型不限			★★
抓取目的:全文文本数字转换;文本印刷时代:Modern	被抓取文本的数量和类型不限		★★	★
抓取目的:全文文本数字转换;文本印刷时代:Historic	被抓取的文本图像数量 < 1000;被抓取文本类型:Simple[注 4]		★★	★
抓取目的:全文文本数字转换;文本印刷时代:Historic	被抓取的文本图像数量 < 1000;被抓取文本类型:noisy[注 5]		★	★★
抓取目的:全文文本数字转换;文本印刷时代:Historic	被抓取的文本图像数量 < 1000;被抓取文本类型:Complex[注 6]		★	★★
抓取目的:全文文本数字转换;文本印刷时代:Historic	被抓取的文本图像数量 > 10000;被抓取文本类型:Simple		★★	★
抓取目的:全文文本数字转换;文本印刷时代:Historic	被抓取的文本图像数量 > 10000;被抓取文本类型:noisy	[注 7]	★	★

续表

目的	被抓取的文本数量和类型	OCR识别且无需校对	OCR识别且需要校对	重新键入
抓取目的:全文文本数字转换;文本印刷时代:Historic	被抓取的文本图像数量>10000;被抓取文本类型:Complex	[注8]	★	★★

注:[1]"★"表示有效方法;"★★"表示从成本节约和准确性综合角度考虑,最有效方法。

[2]Modern:1950年之后印刷的书籍或杂志文本,内容印刷特征以黑白色为主,带有灰色或彩色模式。

[3]Historic:1900年之前印刷的书籍或杂志文本,内容印刷特征以黑白色为主,带有灰色模式。

[4]Simple:印刷非常清晰的单栏目文本,不带有科学符号,只有一种语言,无小字体,无特殊字符、图表或说明。

[5]Noisy:由于灰尘、破缝、退色、涂改、起皱等原因导致文本不清晰、不整洁,其他特征与Simple相同。

[6]Complex:印刷非常清晰的文本,但至少包括下述一个因素:多栏目、带有科学符号、多语种、小字体、特殊字符、图表。

[7,8]仅用于自动做索引时可考虑的文本抓取方法。

2.4 优化数字化生产流程

生产流程是实现某一特定目标而采取的行动的逻辑集合。最大限度地利用现有资源,提高生产效率,无疑会降低单件文献的数字化成本。优化生产流程是在一定时间和成本支出内获得最大产出的一种方法。

生产流程的重要性在于:人力成本通常是最高的成本;障碍和瓶颈不仅消耗资金,而且也浪费时间,资金和时间都是成本的重要组成要素;生产流程是实现项目计划的保障;合理的生产流程可以使风险发生之前就被有效识别,从而避免重大损失,还是产品质量的保障,降低次品的手段。

在优化数字化生产流程时,要考虑的因素有:

(1)分析确定数字化生产过程中需要实施的最重要任务,从而定义关键路径。关键路径应是花费资源最多的工序,如果其他工序占用了过多资源,就要重新考虑这个生产流程是否为最优化。

(2)发生顺序:有时通过简单的次序调整,使任务完成最符合成本节约的宗旨。

(3)并行操作:确定哪些操作可以同时进行,以便设计出最节省时间的生产流程。

(4)确定真正所需的输入输出要素,抛弃不是必需的要素,节省生产成本。

2.5 强化质量管理

质量管理贯穿于数字化项目生命周期的整个阶段,是降低成本的重要方法。质量管理不纯粹是对最终产品的检查,因为任何一个项目不可能有足够的资源来检查每件最终产品,确保零差错。

质量管理的目的在于在项目运行、工作流程和员工操作过程中发现问题。系统误差可以通过修改系统设置来纠正。对于人为错误,可以通过重新培训或调换更适合的员工来处理。

质量管理的重点在于对生产过程的不断完善和持续优化。任何质量管理系统都不可能简单地生成一份包含固定差错的清单。发现差错,就要对工作流程进行调整,使类似差错不再出现。重新数字化对项目而言成本昂贵,质量管理的底线是尽可能减少重新扫描的数量。

参考文献

1 NINCH. The Price of Digitization: Resources. [2006-05-29]. http://www.ninch.org/forum/price.resources.html

2 RLG. Worksheet for Estimating Digital Reformatting Costs: A guide to the preparation of abudget for digitisation. [2006-06-10]. http://www.rlg.org/en/pdfs/RlGWorksheet.pdf

3,8 Simon Tanner. Cost Reduction in Digtisation. [2006-06-20]. http://www.minervaeurope.org/publications/CostReductioninDigitsation-v1-0606.pdf

4 ASSY SA 4digitalbooks. [2006-06-20]. http://www.4digitalbooks.com/digitizing-line.htm

5 kirtas Technologies Inc. Kirtas. [2006-06-23]. http://www.kirtas-tech.com/products.asp

6 ATLZ BookDrive. Automatic Book Scanner. [2006-06-11]. http://www.atiz.com/bookdrive.php

7 Entlich, R. Handwriting recognition for historical documents. [2006-06-12]. http://www.rlg.ac.uk/preserv/diginews/diginews8-1.html

臧国全 武汉大学博士后,郑州大学教授。通信地址:郑州大学信息管理系。邮编450052。

(来稿时间:2006-07-24)