图书馆论坛
Library Tribune
ISSN 1002-1167,CN 44-1306/G2

《图书馆论坛》网络首发论文

题目：家谱知识服务平台众包模式的设计与实现
作者：刘倩倩，夏翠娟
收稿日期：2019-12-01
网络首发日期：2019-12-16
引用格式：刘倩倩，夏翠娟．家谱知识服务平台众包模式的设计与实现[J/OL]．图书馆论坛. http://kns.cnki.net/kcms/detail/44.1306.G2.20191214.1142.004.html

*本文系本文系社会科学基金重大项目“编纂《1949 年以来中国家谱总目》”（编号：18ZDA329）研究成果。

家谱知识服务平台众包模式的设计与实现*

刘倩倩，夏翠娟

摘　要：上海图书馆需要长期征集和数据化家谱，针对家谱文献量大、数据化成本高、耗时长等挑战，在家谱知识服务平台建设中引入众包理念，开发了上传家谱、在线识谱、在线修谱等功能。文章从功能设计、业务流程设计、质量控制、用户激励等方面研究该平台的众包模式。

关键词：众包，家谱，数据化，上传家谱，在线识谱，在线修谱

引用本文格式：刘倩倩，夏翠娟.家谱知识服务平台众包模式的设计与实现[J].图书馆论坛，2020

Design and Implementation of Crowdsourcing Model for Chinese Genealogy Knowledge Service Platform

LIU Qianqian, XIA Cuijuan

Abstract:：Shanghai Library needs to collect and digitize genealogies for a long time. In view of the challenges such as the large number of genealogy documents, high cost of data and long time-consuming, it introduces the concept of crowdsourcing in the construction of Chinese Genealogical Knowledge Service Platform, and develops the functions of uploading genealogies, manual transcription of genealogies and the online editing of genealogies. This paper studies the crowdsourcing model of the platform from the aspects of function design, operation flow design, quality control and user motivation, etc.

Keywords: crowdsourcing, genealogy, data Processing, uploading Genealogy, manual transcription of genealogy Online, online Genealogy Editing

0 引言

上海图书馆（以下简称“上图”）馆藏家谱文献 3 万多种，近年不断将家谱数字化，并探索用关联数据技术揭示家谱中的知识，2015 年推出家谱知识服务平台[1]。要更好地支持家谱利用和研究，还需解决两个问题：一是持续收集散落在各处或新修的家谱，但大范围全面征集较困难；二是对家谱知识的揭示需要将数字化文献数据化。家谱数量庞大、编排繁杂，特别是一些手写字体的老谱，字迹模糊，无法完全借助 OCR 进行全文识别，如果依赖人工识别，成本高，也对工作人员有较高的专业知识要求，若单靠图书馆专业人士，很难快速地全面数据化。为此，上图家谱知识服务平台二期项目引入众包理念，开发上传家谱、在线修谱和在线识谱功能，实现了以人为中心的家谱知识服务。

1 众包的理念与方法

“众包”（crowd sourcing）概念由美国《连线》杂志记者杰夫·豪（Jeff Howe）于 2006 年提出，并没有统一的定义。杰夫·豪将其定义为：“一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的（而且通常是大型的）大众网络的做法。”[2]维基百科对众包的定义是：志愿者或业余人士利用空闲时间解决问题或提出各自观点的做法[3]。BRABHAM 将众包描述为：一种企业在线发布问题，大众群体提供解决方案，赢者获取报酬，且其知识成果归企业所有，以在线分布式的问题解决模式和生产模式[4]。综上，众包的核心理念是充分利用公众力量和智慧来解决较复杂的问题。网络技术发展使众包有广阔的发展前景，基于网络，它能聚集不同背景的人，支持他们在不同地点从不同任务节点执行相同目标的任务，提高工作效率，降低工作成本。

图书馆不以盈利为目的，遵循国际开放标准，与众包模式结合具有天然优势。国外图书馆界有一些成功的众包应用案例。纽约公共图书馆利用众包模式，将馆藏 4 万多份菜单图片转换为可供检索的文本；美国国会图书馆向弗雷克社区（Flickr）发布 3,115 幅没有版

权的照片，获得众多的分类、标注和评论；澳大利亚国家图书馆从海内外 50 多家文化遗产机构收集数字图片，不到 4 年获得 5 万多张[5]。上图对众包模式进行探索：开发了专门的历史文献众包平台，通过完整的任务发布、用户认领、任务执行与协作、任务回收评审的众包流程，达到对馆藏手写资料进行文本化和标引的目的；通过在上图官网验证码中嵌入需人工识别的汉字，以隐形方式引导公众参与创造；组织各种数据竞赛，以活动形式实现众包。这些尝试为众包实践积累了宝贵的经验。

家谱领域面临难以大范围收集、大规模扫描、多体例整理家谱文献及数据化工作量太大等问题，众包模式在其他领域的成功实践带来了启发。国外三大商业家谱网站——Ancestry、Family Search、My Heritage 以为用户寻亲思想，吸引用户上传或创建自己的家谱（需付费），就是众包模式在家谱收集方面的先驱；国内也有中华寻根网、百姓网等以众包思路征集家谱。但从使用情况看，用户上传的家谱格式多样，质量参差不齐，很多用户在网站上新修的家谱仅仅是娱乐，不能考证其真实性，也很难和已有家谱文献产生任何联系，更不能依此溯源。家谱领域的众包需要思考：对基于众包模式收集家谱，如何对公众提交的内容进行质量控制？吸引用户娱乐性地修谱与基于收藏研究目的的收集家谱信息如何平衡？是否可以考虑将已有家谱文献以众包模式进行数据化，以更好地支持用户寻根溯源和学术研究？在数据化众包任务中，如何吸引和留住参与者？如何确保参与者按要求完成任务？如何管理和保存用户贡献的知识和内容？

2 上图家谱知识服务平台建设

上图 2015 年底建设家谱知识服务平台，利用数字人文方法和关联数据技术对《中国家谱总目》和馆藏家谱编目数据进行重组、丰富和格式转换，使以文献为中心的查询检索尝试向数据服务和知识服务转型，希望通过该平台实现全网域范围的家谱文献书目控制，并满足从基于有限已知信息的寻根问祖，到面向特定研究主题的知识发现，再到基于用户贡献内容的知识进化等多层次的用户使用需求。与传统家谱数据库相比，这种建设理念和功能设计可以使资源与用户更好地连接起来，发挥更大的价值。该平台 2016 年正式推出时，在功能上已实现对 597 家家谱收藏机构的 5.4 万多种家谱文献基于概念匹配的检索，以及基于人、地、机构、时间、地点之间关联关系的可视化浏览；提供用户交流和互动的平台，支持用户留言反馈，经过认证的专家登录后还可以对已有数据进行修正和补充，经审核通过后发布在网页上，同时系统会记录用户的每一次修改信息[6-7]。平台推出后受到广泛关注。从用户使用体验和反馈问题看，在满足用户使用需求方面需要进一步努力。比如，并不是每个用户都可以在平台上查找到想要的家谱，这虽然是家谱文献流传和保存的客观问题，但无疑会打击用户的使用热情；而家谱资源收集、用户体验、功能设计、系统维护和个性化服务上均有提升空间[8]。基于对家谱知识服务平台的优化需求，以及引入众包模式满足用户多方面需求考虑，上图 2017 年启动家谱知识服务平台二期项目，在进一步优化系统的基础上，增加新功能——上传家谱、在线识谱、在线修谱。

3 上图家谱知识服务平台众包功能的设计与实现

上图家谱知识服务平台以用户使用需求为导向设计，众包任务的成败关键在于用户。图书馆是非盈利性的公共服务机构，上图在家谱收藏方面具有较大影响力，一期家谱知识服务平台聚集了较多用户，很多用户愿意自发参与平台建设，贡献自己的力量，这为众包模式的实现提供了可能。

上图家谱知识服务平台用户分为两种：普通用户和有专业研究需求的用户（对家谱信息感兴趣的专家学者、民间团体等）。普通用户一般是已知某先祖名人的姓名或家族的居地、堂号等，希望通过平台找到家谱文献，即寻根问祖。普通用户若在平台上找到跟自己有关

的家谱时，可能希望进一步续修或将已经续修的家谱上传，在平台上得以永久保存续谱。如果用户无法找到想要的家谱时（事实上，这是多数人会遇到的客观问题），可能希望通过便利的方式新修家谱，即在线修谱。专业用户对平台的需求除查找感兴趣的家谱文献外，还希望与图书馆、其他家谱专家针对收藏情况、家谱专业研究问题等进行交流互动，他们对某一姓、某个家族的家谱了解深入，可能已搜集了一些图书馆尚未收藏的有价值的家谱文献，如果平台支持将家谱直接上传，支持用户基于图书馆的家谱文献通过各种方式进行公共家谱知识生成，会进一步提高用户对平台的使用意愿，增强平台的用户黏度。

根据平台建设理念和用户需求，对家谱众包任务进行设计，任务包括：（1）现有家谱资源的完善和平台优化，即用户对家谱文献进行反馈、纠错，对平台建设提议（平台一期建设已实现该功能）；（2）新家谱资源的产生与搜集，即上传家谱和在线修谱/续谱；（3）图片形式的家谱文本化，因世系图是家谱中最重要的模块，暂只设计将上图开放的 8,000 多种家谱中的世系图作为众包任务之一。基于以上考虑，上图家谱知识服务平台二期项目设计了 3 个众包的功能：上传家谱、在线识谱（世系图模块）、在线修谱（续谱）。

3.1 众包功能与业务流程设计

3.1.1 上传家谱

上传家谱支持登录用户将已有的家谱上传并保存在网站上。除支持用户上传家谱的全文（word、pdf、zip 等格式）外，还设计了元数据方案，支持用户按要求填写内容后自动生成和保存元数据。上传家谱等同于电子版的家谱捐赠，是对上图普通家谱捐赠渠道的补充，不仅避免传统捐赠的弊端，如降低捐赠人的时间经济成本，而且自动生成元数据功能，可使资源得以更方便地保存和检索。上传家谱的业务流程分为平台用户和管理员两条线。

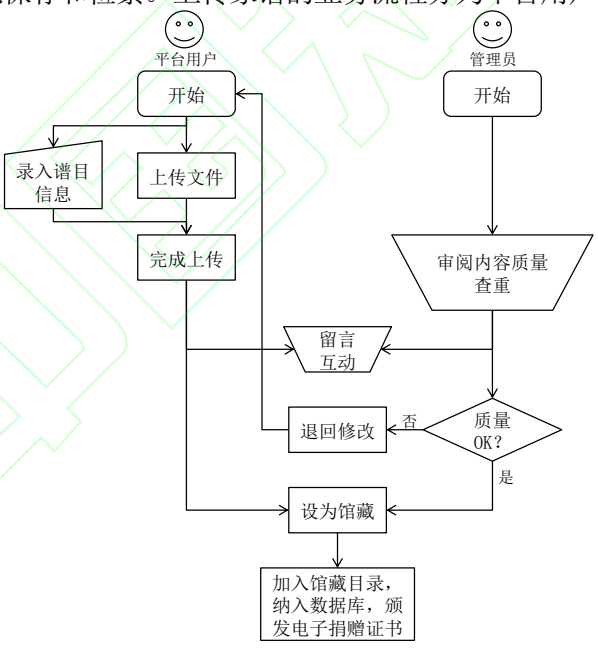

图 1 上传家谱业务流程图

如图 1 所示，用户登录后，可使用上传家谱功能，录入平台提示的谱目信息和个人信息，将符合格式要求的文档上传到平台上。系统管理员则对用户上传的文件和录入的信息进行审阅，进行查重和质量校验。如果馆内现存文献中无相同家谱，且质量符合馆藏要求，管理员将其设为馆藏，用户填写的各字段信息作为元数据保存在后台数据库，并支持在前台网页中检索；未设为馆藏的，系统会自动保存所有信息，上传用户可随时查看，管理员也可要求用户继续对上传的家谱数据进行修改和完善，直至符合馆藏要求为止。

要求用户上传家谱时必填的字段就是支持检索的元数据信息，如谱籍地、书名项、责任者项、版本项中的版本年代、载体形态项、装订项、收藏者项，见图 2 中带*号的字段。用户上传的文件有格式要求，如封面仅支持 jpg、jpeg、png 格式，全文仅支持 doc、docx、pdf、zip 格式，数码图像精度不能低于 300DPI 等。

图 2 上传家谱功能页

3.1.2 在线识谱

在线识谱是针对平台已外网开放的 8,000 多种家谱资源，邀请用户协同图书馆完成数据化加工，目前众包任务主要是对世系图进行在线抄录。世系图文本化后，可支持对世系图中某个人名的检索，并在此基础上依据数据人文方法发现更多的研究线索，如对人物的统计分析中发现知识，从人物之间的关系发现更多的故事。家谱中的世系图往往篇幅较大，且各种连线复杂，抄录需要较长时间，对利用碎片化时间参与众包任务的用户来说，很难确保完成度。所以，在线识谱模块的众包任务虽以每种家谱的世系图为最小单位，但引入多角色协作任务的功能，同一个任务支持多人共同编辑完成。

平台将除系统管理员之外的用户分为两种：专家用户和普通用户。每种用户的权限各不相同，系统管理员权限最大，负责众包任务的发布（开放家谱）和对其他用户的管理维护；专家用户是上图定向招募和认证的家谱领域专家，可以在平台上认领众包任务，自己完成任务，也可以邀请其他专家用户和普通用户协同，认领任务的用户对整个任务的进度和质量负责。普通用户作为协同者参与到众包任务中。

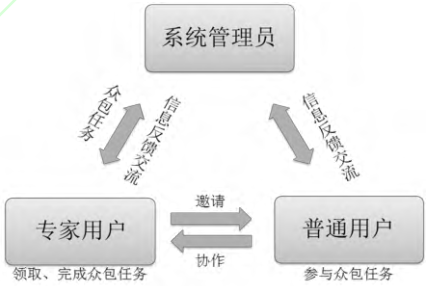

图 3 在线识谱的用户组织

每个家谱的世系图识别为一个单独的任务，在多角色用户协同下共同完成。在这种业务流程模式下，不同类型用户的任务界面也是不同的，对每个用户来说，只有被自己认领的家谱才能被修改，或者作为被邀请的协作者才能编辑世系图。

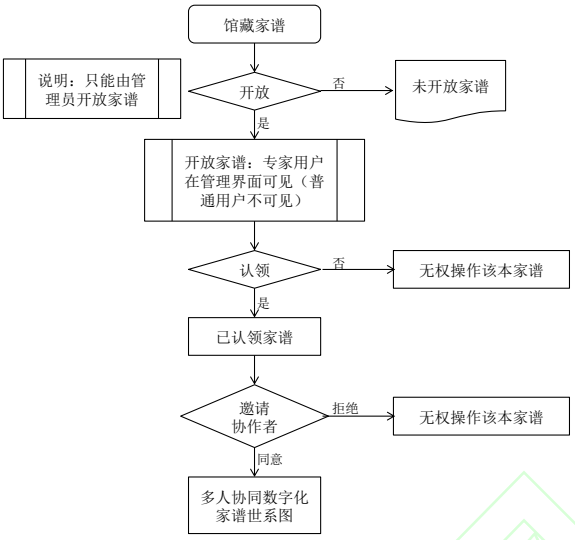

图 4 在线识谱业务流程图

图 5 专家用户在线识谱管理界面

3.1.3 在线修谱

在线修谱功能支持对新修或续修家谱感兴趣的用户从零开始修谱。该功能集成在时光公司的有谱平台上，可实现谱目的快速录入、家谱内容按模块完整录入、机构人员管理、家谱出版等功能。也支持邀请和添加协作人员，共同完成家谱的修撰。修谱前，需要新建机构和添加机构所属人员，才能编辑谱目。整体业务流程见图 6。用户可以按照系统提示一步步创建家谱。在系统上进行在线修谱或续谱操作，非常便利，比如世系表的编辑便捷、人员协作管理方便。用户如有需要，还可在创建好的家谱基础上出版纸质版。

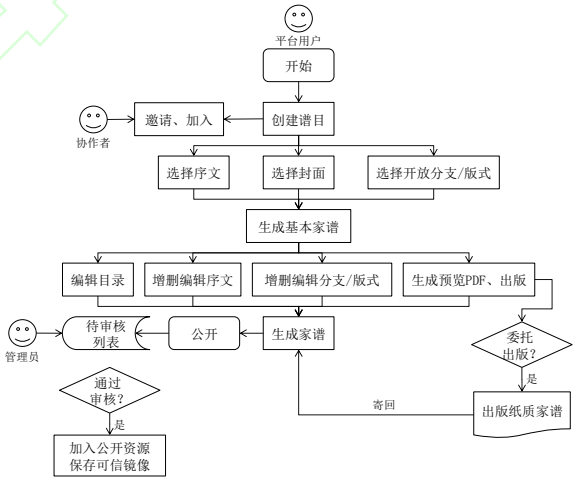

图 6 在线修谱业务流程图

3.2 家谱众包的质量控制

众包项目的完成需要依托广大网民的参与和支持，由于参与者的知识结构、社会背景等不同，加之对众包任务的理解可能有所不同，这对众包任务的完成质量都有影响。众包的质量控制成为项目需考虑的关键问题。家谱众包项目的质量控制从任务开始前、任务过程中和任务完成后 3 个阶段的 3 个层面进行设计（图 7），确保对众包的任务结果进行反复的确认和修正，实现对家谱众包项目的全过程监管，以保证其准确性与可信度。

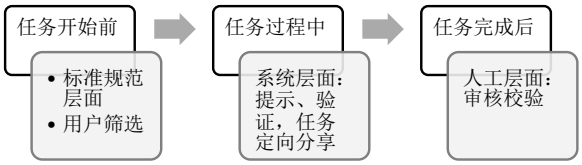

图 7 家谱众包的质量控制过程

任务开始前，主要从标准的层面对家谱众包的任务进行规范。比如，在上传家谱的功能设计上，上图请馆内家谱专业研究工作人员共同讨论著录标准，设计了相应的元数据方案，规范了用户上传家谱时必填的元数据信息；在线识谱和在线修谱功能中也有类似的规范字段要求。另外，在线识谱还对参与用户进行了筛选控制，在已有用户群体招募核心用户，邀请他们作为专家用户参与对感兴趣家谱的众包任务，在平台的用户角色中对其进行认证，并给予更高的责任和权限。

任务过程中，在系统层面会对用户录入的内容进行提示说明和验证。比如，在上传家谱填写元数据信息时（见图 8），鼠标移至录入框，上方会出现黑底白色的录入示例；用户保存时，系统会对录入的内容进行验证，若未完成必填字段的录入，则不允许保存。此外，在线识谱任务难度高耗时长，依赖一个用户难以保证任务的结果质量，系统设计上支持任务的定向分享，采用“专家用户+普通用户”的合作模式，多角色协作完成众包任务。

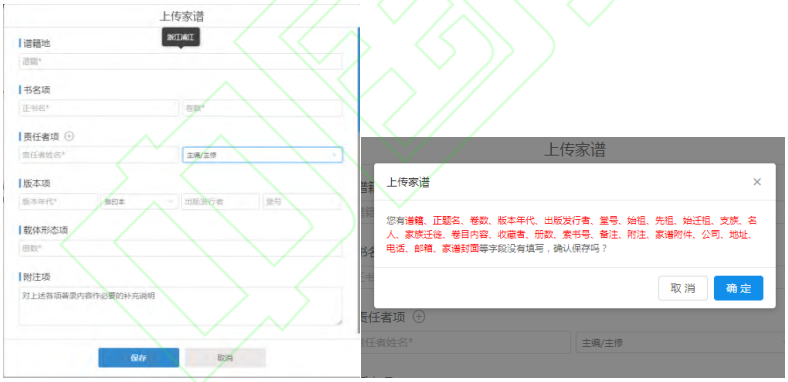

图 8 系统层面的提示与验证

任务完成后，对众包任务进行人工层面的审核与校验。比如，上传家谱中，平台的系统管理员和家谱研究专家对上传的家谱进行人工审核，如发现家谱质量不符合要求，可以将家谱退回至用户，要求修改。

3.3 用户激励策略

成功的众包需要用户的积极参与，而了解用户参与众包的动机是制定有效激励机制的前提。用户参与众包的动机分为内部动机和外部动机，外部动机包括报酬激励和社会性动机（如反馈、认可、奖励、归属感），内部动机包括基于自我享受的动机（如研究兴趣、内在满足感和成就感）、基于社区的动机（参与感、分享等）[9]。上图是非盈利组织，家谱众包的资金激励不足，对众包参与用户的外在动因激励主要是非物质形式的精神奖励，如用

户上传的家谱经工作人员审核，决定纳入馆藏的，颁发电子版捐赠证书，与线下捐赠家谱颁发的捐赠证书具有同等效力，受到捐赠用户的重视；用户在平台上纠错或反馈问题，系统管理员给予积极回应和支持，也有利于提高众包用户的参与积极性。

除外在动因，家谱众包主题也在一定程度上有利于家谱爱好者的内在动因实现。比如，馆藏数量众多的家谱文献支持参与用户的研究兴趣，用户在参与众包的过程中对文献更为熟悉，满足研究需要；对专家用户的角色认证是对活跃用户的肯定，有利于众包用户在参与过程中提升自我肯定和成就感；在线识谱多角色协作模式则是基于用户的社交参与需求，利用专家用户的核心影响力来组织更多的用户参与到众包任务中。

4 用户社区的构建

除以上家谱众包任务外，家谱知识服务平台重视用户的经验交流、信息与知识的分享讨论。平台一期项目在平台上建立了家谱交流反馈模块，支持用户留言反馈，对家谱文献进行纠错。为了达到宣传和交流目的，平台用户建立了 QQ 群——上图家谱核心用户群。QQ 作为熟知的社会化软件，支持创作、分享内容，可提高用户的共享积极性，实现众包参与者的知识共享、工作经验交流，也有可能使一些尚未参与平台众包的目标用户对平台有所了解，并转化为家谱知识平台的贡献者。上图专门派工作人员对群进行维护，有意识地将其打造为一个对平台众包功能进行补充的社区，在社区内实现众包用户与非众包用户的交流和知识共享。

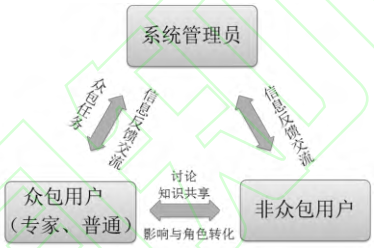

图 9　用户社区的互动交流

5 总结与展望

家谱知识服务平台二期项目于 2018 年投入使用，实现上传家谱、在线识谱和在线续谱等众包功能，吸引了用户的关注，使得平台在上图的所有系统中的浏览量排名靠前。众包作为创新的工作方法，在家谱系统中应用，使收集家谱文献更便捷，“专家用户+普通用户”协同工作模式为家谱文献数据化工作提供了新方法，在未来大规模的家谱数字化处理工作中应有广泛的应用前景。

上图家谱知识服务平台也存在需优化的地方，主要有 4 个方面。

（1）加大宣传推广。家谱领域专业，用户属于小众群体，因此要加大宣传力度，扩大用户群体，使更多用户参与家谱资源的利用和知识创造中来。

（2）用户激励机制。用户参与是众包任务成败的关键，目前家谱众包项目的资金激励不足，精神激励有待丰富。家谱众包任务专业，数量多难度大，需设计更有效的组合激励策略，激发用户参与兴趣，如引入积分和等级体系，任务设计吸取游戏化设计理念。

（3）提升使用体验。用户使用体验直接影响使用的意愿，可优化现有界面，在平台中嵌入各种工具，如支持上传家谱图片的在线处理工具，辅助识谱的 OCR 工具（将工具 OCR 与人工抄录结合），用户操作更方便。

（4）完善用户社区。用户社区是平台重要的配套性设施，目前平台交流反馈功能简单，主要依靠外部 QQ 群交流，未来可在平台上集成功能更强大的用户互动社区，使用户在平台上参与度更高。

参考文献

[1] 中国家谱知识服务平台[DB/OL]. http://jiapu.library.sh.cn, 2019-08-20.

[2] Howe Jeff．The rise of crowdsourcing［J］．Wired，2006，June: 20．

[3] WIKIPEDIA A．Crowdsourcing［DB/OL］．http://en.Wikipedia.org/wiki /Crowdsourcing, 2019-08-20．

[4] BRABHAM D C．Crowdsourcing as amodel for problem solving: an introduction and cases［J］． The International Journal of Research into New Media Technololgies，2008，14(1) : 75－90．

[5] 盛芳, 李正龙, 焦坤,等.众包与众包馆员制度：助推图书馆服务转型[J]. 图书情报知识, 2012(4):95-102.

[6] 夏翠娟, 刘炜, 陈涛, 张磊.家谱关联数据服务平台的开发实践[J]. 中国图书馆学报, 2016(3).

[7] 夏翠娟, 张磊.关联数据在家谱数字人文服务中的应用[J]. 图书馆杂志, 2016(10):26-34.

[8] 赵雪芹, 邢慧.家谱知识服务平台用户持续使用意愿研究——以上海图书馆家谱知识服务平台为例[J]. 图书馆, 2019(3).

[9] Kaufmann, N., Schulze, T., Veit, D. More than fun and money. Worker Motivation in Crowdsourcing - A Study on Mechanical Turk［EB /OL］．https://pdfs.semanticscholar.org/ff27/9098481a87faa4498fa9088cd9bd835e9a3e.pdf,2019-08-20．

作者简介:刘倩倩（通信作者，ORCID：0000-0002-8111-5154，qqliu@libnet.sh.cn），上海图书馆系统网络中心助理工程师；夏翠娟，上海图书馆系统网络中心高级工程师。

收稿日期：2019-12-01

（责任编辑：沈丽霞）