

国内外开放数据竞赛的案例分析及运行机制探索*

赵宇翔¹ 仝 冲¹ 张轩慧² 张 磊³

(1 南京理工大学经济管理学院 2 南京大学信息管理学院 3 上海图书馆)

摘 要 本文对现有的开放数据竞赛的文献进行系统梳理, 结合4个国内外不同类型的开放数据竞赛案例, 对开放数据竞赛的特点进行分析讨论。在案例分析的基础上总结出5种开放数据的特征和6种开放数据平台机制, 提炼出开放数据竞赛的基本框架, 并针对框架对开放数据竞赛的关键流程进行阐述, 最后对我国开展开放数据竞赛提出若干建议。

关键词 开放数据竞赛 价值共创 开放创新 运行机制 案例分析

DOI: 10.13663/j.cnki.lj.2020.03.015

Case Analysis and Operational Mechanism Exploration of Open Data Competitions at Home and Broad

Zhao Yuxiang¹, Tong Chong¹, Zhang Xuanhui², Zhang Lei³

(1 School of Economics and Management, Nanjing University of Science and Technology; 2 School of Information Management, Nanjing University; 3 Shanghai Library)

Abstract The study first reviewed the existing literature on open data competitions and identified their characteristics by investigating four different types of open data competition cases at home and abroad. Accordingly, it summarized the characteristics of open data and open data platforms, and proposed the basic framework of open data competitions. Based on the framework, the key procedures of open data competitions were presented. Finally, it put forward suggestions to provide the necessary guidance for domestic open data competition organizers.

Keywords Open data competition, Value co-creation, Open innovation, Operational mechanism, Case analysis

0 引言

随着互联网的发展, 越来越多的政府机构和企业组织开始对数据进行公开, 数据开放成为大势所趋。为了更好地使用和利用企业和政府提供的开放数据, 激励更多的公众参与到该过程中, 部分机构和平台发起了各类开放数据竞赛(open data competition)活动。开放数据竞赛是主办方以推动互联网公共开放数据的挖掘和使用为目的, 旨在收集某一领域对网络开放数据成功使用案例的一种竞赛形式^[1]。开放数据竞赛与开源软件竞赛和众包竞赛在形式上有一定的相似性, 均采用竞赛形式吸引更多的参与者, 激发新的服务创意和产品原型

的开发, 最终实现更大程度的开放创新(open innovation)。开放创新是一种提倡企业可以并且应当同时使用外部和内部创新或想法的范式^[2]。该范式是一种有利于促进服务和产品的制造者和使用者共同创造并实现产品和服务的价值共创(value co-creation)模式^[3]。价值共创模式对于组织者和参与者来说都具有重要

* 本文系国家社会科学基金重大项目“面向知识创新服务的数据科学理论与方法研究”(项目编号: 16ZDA224)、国家自然科学基金项目“面向数字人文研究的图书馆开放数据体系构建与服务模式设计”(项目编号: 18BTQ027)和中央高校基本科研业务费专项资金资助(项目编号: 30919011104)的研究成果之一。

意义。对组织者而言,价值共创可以降低成本及提供更好的产品和服务。对参与者而言,价值共创可以提高参与积极性,最终改善用户体验。在开放数据竞赛中,数据提供方和参赛者通过开放数据提出问题并开展知识创新服务,从而实现基于数据驱动的价值共创。

开放数据竞赛模式既有利于开放数据提供方寻求解决问题的方案,也有利于参赛人员锻炼自身能力,提高在专业领域的知名度。目前在国内外都已经开展了一些实践探索,越来越多的政府机构和企业组织逐渐关注这一活动形式。然而,针对一些科学研究问题,譬如,如何更有效地组织开放数据竞赛?如何吸引更多的参赛者融入开放数据竞赛的各个环节?如何吸引社会各界力量介入开放数据竞赛活动?目前国内外还鲜有研究给出答案。本研究将对现有的开放数据竞赛的案例进行分析总结,提炼开放数据竞赛的基本特征,并在此基础上提出开放数据竞赛的框架和运作机制,旨在为我国开放数据竞赛的规划和设计提供初步的理论参考和实践指导。

1 开放数据竞赛研究综述

国外学者率先针对开放数据竞赛的议题进行了理论和实践研究。Hjalmarsson等提出开放数据竞赛通常指在规定的时间内,参赛者利用开放数据开发出新的想法和原型的一种竞赛形式,它逐渐成为刺激新产品开发的流行趋势^[4-5]。这种参与形式利用了竞争的特点和一些激励因素(例如物质奖励、名声以及经验等)刺激公众开发更高性能的软件应用^[6-7]。从竞赛主办方的角度,研究者普遍关注竞赛的设计问题。Schöttner等对不同的开放数据竞赛奖项设置进行了对比研究,结果表明竞拍制和排名制的奖项设置方式分别会吸引不同类型的参赛者^[8]。Brunswick等以开放数据竞赛为例探究开放创新竞赛在举办过程中的透明度对结果的影响,研究发现虽然部分学者认为提高透明度可能会产生过度复制的问题从而影响参赛结果,但是整体看来提高透明度有利于产生更好的参赛作品^[9]。从参赛者的角度,现有的研究大多关注参赛者的行为。Bullinger等发现具有

较高和较低合作程度的参赛团队会在开放数据竞赛中取得较好的绩效,而合作程度中等的参赛团队在开放数据竞赛中的表现较差^[10]。Huang发现开放数据竞赛的形式可以激发具有更高抱负的人的自我选择,而缺乏自我追求的人则会主动规避开放竞赛^[11]。Füller等对参赛者的参赛行为进行分类,并从交流和贡献层面确定了6种不同类型的参赛者:社交者、创意提供者、领导者、高效参与者、被动创意提供者以及被动评论者^[12]。从参赛作品角度,部分学者关注开放数据竞赛的作品向开放数据市场的转化问题。Hansen等对开放数据竞赛的作品是否可以产生持续导向型的创新进行探究,并发现新人获奖者的作品往往有更好的投资收益^[13]。此外,部分学者探究了包含开放数据竞赛在内的开放创新模式对价值共创机制的贡献。Wang等基于开放创新过程中多方主体的协作模式构建了相应的价值共创模型^[14]。Piller以艺术家电商在线社区“Threadless”为例探究开放创新竞赛中参与者与生产者如何更好地实现价值共创^[15]。

相比之下,国内对于开放数据竞赛的探究起步较晚,且主要集中在开放数据竞赛的实践活动上。目前国内较为知名的开放数据竞赛有上海开放数据应用大赛SODA以及上海图书馆开放数据开发应用大赛。前者是以提供开放数据集下载的形式让开发者参与到开放数据的数据分析过程中,更好地解决上海的城市问题^[16];后者则是通过上海图书馆搭建的开放数据平台,以应用程序接口(API)的形式允许开发者在个人开发的产品中获取上海图书馆的开放数据,从而实现对开放知识服务模式的新尝试^[17]。然而,目前国内学界对于开放数据竞赛的理论探索,尤其是开放数据竞赛的模式规划和设计,以及开放数据竞赛的运作机制等问题还鲜有研究。

总的来说,开放数据竞赛作为一种新兴的开放创新模式,正逐渐受到国内外学者的广泛关注。国外的开放数据竞赛起步较早,相应研究也较为成熟,对开放数据竞赛的形式和机制也有了初步探索,但这些探索多聚焦在微观层面,尚缺乏对现有案例的分析总结以及完整清

晰的运行机制和框架，因此无法满足国内开放数据竞赛活动的实践指导需求。本文主要借助具体的开放数据竞赛案例，对开放数据竞赛的框架和运行机制深入挖掘，从而为我国开放数据竞赛项目的推广和发展提供参考建议。

2 国内外开放数据竞赛案例分析

开放数据竞赛是基于开放数据开展的知识创新服务。迄今为止国内外已有较多地对开放创新形式的探索和应用，包含众包^[18-20]、开源软件^[21-23]、维基^[24-26]等，本文尝试归纳了开放数据竞赛与其他开放创新形式的对比（参见表1），从中可以看出开放数据竞赛与其他形式的开放创新存在很多异同点。开放数据竞赛可以

看作是建立在众包、开源软件等机制基础上结合开放数据使用的创新，由于这一形式目前还处在探索阶段，因此部分开放数据竞赛依然带有众包等开放创新形式的特点。通过对不同开放创新形式进行对比，笔者发现开放数据竞赛的独特性在于其依托于开放数据平台、参与形式多为团队参赛以及同时具备较强的竞争性和合作性。

本研究选取了国内外不同的4个开放数据竞赛进行案例研究，这4个竞赛在开放数据获取、参赛流程设计、奖项设置等方面都较为成熟完善，且在相应的领域都有较大的知名度，具有一定的代表性。笔者结合4个案例的网站和资料信息，对案例进行总结分析和讨论。

表1 开放创新形式对比

	开放数据竞赛	众 包	开源软件	维 基
平台	开放数据平台	众包网站	开源平台	无统一平台
主体	参赛者、主办方等	参与者、平台方、任务发布方等	参与者	参与者
目标	提交数据分析方案或原型开发作品	完成发布方发布的众包任务	开发、修改、整合、公开开源软件	创建、修改、删除词条
参与者类型	数据科学等相关学科专业人员	一般公众	开发者	相关领域专业人员
参与形式	团队为主	个体为主	个体	个体
是否有报酬	有	有/无	无	无
竞争性	强，且存在与团队之间	较强，且存在与个体之间	弱	无
合作性	强，且存在是团队内部	一般	强	强
创新性	在设计作品的过程中较强	在创新型众包项目中较强，其他形式的众包项目较弱	较强	较弱

2.1 案例概述

在过去的10年中，越来越多的政府和机构开始向公众开放社会重要领域的数据集，例如教育、健康医疗、消费、衣食住行等方面。这些曾经只在政府和机构内部获取和使用的数据，如今可以通过互联网在诸多平台上获取^[27]。2009年美国联邦政府发布了data.gov开放数据平台，提供了上千个政府开放数据集以供开放获取^[28]。目前的开放数据竞赛大体可以分为两种：一类是参赛者可以从平台获取开放数据集，采用一定的数据分析和数据挖掘技术解决问题，以国外的Kaggle平台和国内的SODA开放数据竞赛为例^[16, 29]。另一类则是主

办方提供应用程序接口（API）、统一资源标志符（URI）等数据和接口形式，参赛者需要开发App、微站或小程序等信息系统对开放数据进行获取和利用，形成具体的服务或原型，以国外的LinkedUp项目和国内的上海图书馆开放数据开发应用大赛为例^[17]。这两类虽然在开放数据竞赛的形式上有所区别，但本质上都是对开放数据进行更好的开发和利用。

2.1.1 Kaggle 开放数据竞赛

Kaggle成立于2010年，是一个数据建模和数据分析平台^[29-30]。企业可以在Kaggle上发布一些开放数据，同时提出需要实际解决的问题。参赛者可以组队参与，针对某一问题提

出基于开放数据的解决方案,最终由问题提出方选出最佳方案,方案被采纳的参赛者将获得五千到一万美元不等的奖金。除此之外,Kaggle官方每年还会举办一次开放数据竞赛,吸引了广大的数据科学爱好者。Kaggle是基于众包(crowdsourcing)模式的开放数据分析平台,它通过众包的形式招募参赛者参与到开放数据分析和竞赛中。

Kaggle上的竞赛模式有很多种类,例如“Featured”模式面向专业数据科学家,“Research”模式面向一般的数据科学爱好者。总体上看,不同模式下的规则都需要参赛者使用出题方提供的开放数据,建立模型提出解决方案。Kaggle的参赛者大体分为两类:一类是职业参赛者,这类参赛者以物质奖励为目标,追求高奖金和高排名;另一类是业余参赛者,这类参赛者以提升自身能力为目标,追求知识的获取和经验的积累。

2.1.2 LinkedUp Project & Challenge 开放数据项目及竞赛

LinkedUp Project是欧盟推动的开放数据项目,旨在推动网络上可以获取的开放和关联数据的利用,尤其是教育机构和组织^[1]。它通过促进Veni、Vidi和Vici组成的LinkedUp Challenge(开放数据竞赛)来实现这一目标。在开放数据竞赛中参赛者需要开发创造用于分析或集成开放数据的工具,以达到促进教育发展的目的。LinkedUp Challenge从2012年开始到2014年结束,连续举办了Veni、Vidi和Vici三场竞赛。参与者需要完整参与三场竞赛,分别完成产品的设计、开发和运营。Veni竞赛中参赛者需要对开放数据在教育领域的应用提出自己的见解及相应的方案;Vidi竞赛中参赛者需要开发针对教育目的的App和原型,作品可以存在一些bug但是需要能够有可以稳定运行和部署的特征和证明;在Vici竞赛中,参赛者需要开发完整且能稳定使用的产品,并积累一定的用户群体。该项目的参赛者群体是一些关注教育领域的数据科学家,希望通过参加LinkedUp Challenge推广自己的产品和应用。LinkedUp开放数据竞赛的特点在于该竞赛最终需要提交一个完整且投入市场的产品,因此该

竞赛对于开放数据在实际场景中的应用方式有着重要的指导作用,也有利于提高大众对于开放数据竞赛的认知。

2.1.3 SODA 开放数据创新应用大赛

SODA开放数据创新应用大赛是在国家发改委、工信部和中央网信办等国家部委指导下,以充分利用长三角丰富的开放数据资源以及雄厚的大数据产业基础为目的的开放数据竞赛^[16]。该数据竞赛涉及交通、餐饮、城市管理、教育、新零售、旅游、环境保护、医疗、金融和文化等领域的政府开放数据和企业数据。该大赛包括初赛和复赛两阶段,初赛提交作品设计方案、复赛提交作品原型和原型文档。SODA的开放数据主要是通过数据集下载的方式获取,参赛者将数据集中的数据样本直接整合到开发的作品中,因此从技术评估角度来看,注重参赛作品对于数据的清洗、挖掘、关联、可视化和应用整合过程。该竞赛设置了创业基金、获胜团队宣传以及亚马逊应用创新中心孵化等奖项,吸引了众多创业人员的参赛。

2.1.4 上海图书馆开放数据应用开发竞赛

开放数据应用开发竞赛是上海图书馆利用其丰富的特色馆藏资源,并和多家面向知识创新的企业和机构合作,面向大众推出的开放数据竞赛。该开放数据竞赛依托于上海图书馆的开放数据平台,通过关联数据的形式公开上海图书馆的数字人文项目数据。这些关联数据包括基础知识库、文献知识库、本体词表等人文数据以及项目中涉及的多种数据清洗和转换工具、研究成果等^[17, 31]。开放数据平台采用REST API、Sparql Endpoint、内容协商(Content Negotiation)等技术形式向开发人员提供接口,以促进数据的开放获取、共享和使用。该竞赛已经成功举办了三届,产生了众多的参赛作品和优秀的获奖作品,其参赛者多是对数字人文项目和对历史人文知识感兴趣的数据科学爱好者。

综上所述,表2概况了4个开放数据竞赛的基本情况。

2.2 分析与讨论

2.2.1 开放数据对比

开放数据是指免费供个人使用和重新发

表 2 开放数据竞赛案例对比

开放数据竞赛	Kaggle	LinkedUp	SODA	上海图书馆
开放数据平台	Kaggle 开放数据平台	无统一开放数据平台	SODA 开放数据平台	上海图书馆开放数据平台
数据来源	全球政府、企业、机构等	全球教育相关网站	长三角地区政府、企业、机构等	上海图书馆、CADAL 项目管理中心、哈佛大学计量社会学系等
竞赛任务	提出开放数据分析和挖掘模型	开发开放数据应用并投入使用	设计开放数据应用	设计开放数据应用原型
主办方	美国商业公司 Alphabet	Open Knowledge International	上海大数据应用创新中心	上海图书馆
研究领域	全领域	教育领域	全领域	数字人文
竞赛形式	注册账号；选择项目；提交方案	注册报名；连续参与三场比赛	注册报名；参与初赛复赛决赛	邮箱报名；提交作品
评比方式	内部评审	内部评审 + 公开投票	内部评审	内部评审
奖金	由项目发起人自行确定	1 000-2 000 欧元	2 万-10 万元	0.5 万-10 万元
提交形式	问题解决方案	投入市场使用的成熟的应用	应用原型	应用原型

布，且不受版权、专利或其他控制机制限制的数据^[32]。随着数据开放运动在全球范围内的兴起，越来越多的企业和机构开始重视开放数据的价值。在开放数据竞赛中，开放数据的形式和质量也是影响开放数据竞赛的重要指标。高质量的开放数据形式可以促进更高的价值创造^[33]。本文基于夏义堃等^[34]提出的数据源类型、结构类型、整合方式、获取方式和使用方式 5 个方面对上述 4 个案例的开放数据形式进行对比，如表 3 所示。这 5 个方面在宏观上体现了不同的开放数据竞赛在数据源、技术、流程和管理层面的差别。

在 4 个案例中，Kaggle、SODA 和上海图书馆都建立了相应的开放数据平台用于对开放数据资源的整合。LinkedUp 则没有设立自己的开放数据平台，而是要求参赛者在互联网上自行搜寻可用的开放数据。在开放数据的获取方式上，Kaggle 和 SODA 是通过提供开放数据集，让参赛者自行下载使用，上海图书馆则是通过提供多种数据借口供参赛者调用。

(1) 从数据源角度来看，LinkedUp 具有最开放的数据源，但由于并没有对数据本身的质量进行监管，因此无法对数据进行统一管理，其余 3 个案例中的开放数据均经过平台的监

管，有较强的完整性和可用性，其中上海图书馆的开放数据还具有较强的可加工性。

(2) 从技术层面看，开放数据平台提供的数据分析和数据存储等技术是开放数据竞赛值得关注的问题。在 4 个案例中，LinkedUp 并不提供相应的技术，其余平台提供了数据搜寻和数据下载的技术，提高了数据的可获取性和可用性，在此基础上上海图书馆提供了非结构化数据的加工技术，进一步提高了数据的可用性。

(3) 从流程层面看，即数据的生命周期角度，上海图书馆和 SODA 较为重视开放数据的收集流程，而 Kaggle 和 LinkedUp 则降低了开放数据的收集门槛，也一定程度影响开放数据的质量。

(4) 从管理层面来看，不同案例中对于开放数据的边界有不同的定义，在现有的开放数据体系下，完整性指的是所有可开放的数据都应开放，在法律体系的约束下提供所有可用的开放数据集^[35]。LinkedUp 将所有互联网上可使用的数据均作为开放数据源，Kaggle 将所有提交到 Kaggle 平台的数据均视为开放数据，SODA 和上海图书馆通过一定标准选取了平台所用的开放数据，同时鼓励参与者使用平台之外的开放数据。由此可见，除了 LinkedUp 之

外,其他3个开放数据竞赛在开放数据的选取上都有开放数据平台作为依托,Kaggle甚至将竞赛使用的开放数据限制在平台内部。

表3 开放数据形式对比

开放数据竞赛	Kaggle	LinkedUp	SODA	上海图书馆
数据来源类型	平台数据集	互联网公开数据	平台数据集	平台数据集
数据结构类型	结构化和非结构化	结构化和非结构化	结构化	结构化和非结构化
整合方式	任务发布方提交给平台整合	无整合	企业和平台共同整合	平台整合
获取方式	下载	端口或下载	下载	端口
使用方式	分析或挖掘	调用或整合	整合	调用

2.2.2 开放数据竞赛形式对比

为了更好地推进数据开放机制,实现数据价值,激发更多的数据科学爱好者参与到数据开放运动中,国内外的很多政府和企业机构举办了开放数据竞赛。目前的开放数据竞赛主要分类两类:一类是政府主导的开放数据竞赛,旨在推动政务透明和数据公开,并在此基础上寻求解决问题的方案,代表案例有 LinkedUp 和 SODA 的开放数据竞赛。另一类是企业和机构主导的开放数据竞赛,旨在进行算法和模型的改进或推广数据的开放和资源共享使用,例如 Kaggle 和上海图书馆的开放数据竞赛。如表4所示,本文从主办方、项目提供方、开放用户、开放数据平台、参赛方、参赛作品、评审方和投票方几个方面对4个开放数据竞赛的案例进行了对比分析。

表4 开放数据竞赛涉及主体对比

开放数据竞赛	Kaggle	LinkedUp	SODA	上海图书馆
主办方	提供平台、奖金	提供平台、奖金	提供平台、奖金	提供平台、数据、奖金
项目提供方	提出问题和数据并提供奖金	无	无	无
开放用户	向平台提供开放数据	在互联网公开数据	向平台提供开放数据	无
开放数据平台	整合并提供开放数据	无	整合并提供开放数据	整合并提供开放数据
参赛方	数据科学爱好者;求职者;职业参赛者	教育领域的数据科学爱好者	求职者和创业人员	关注数字人文和历史领域数据的科学爱好者
参赛作品	算法、模型等解决方案	投入市场运营的产品	算法、模型等解决方案或者产品原型	产品原型
评审方	评估参赛作品	评估参赛作品	评估参赛作品	评估参赛作品
投票方	无	对作品进行投票	无	无

目前开放数据竞赛还处在兴起阶段,在形式上尚无统一的标准。从表4可以看出,只有主办方、参赛方和评审方是每个开放数据竞赛都存在的主体,构成了开放数据竞赛基本框架最重要的元素。

(1) 主办方负责进行开放数据竞赛的筹办并提供竞赛奖项,其中上海图书馆还负责提供开放数据。

(2) 项目提供方是 Kaggle 特有的主体。Kaggle 是从众包平台衍生出来的开放数据竞赛平台。项目提供方类似于众包中的问题提出

者,负责提出问题和开放数据并提供奖金。在其他的案例中,开放数据竞赛需要参赛者自主提出问题、从开放数据平台获取必需的数据并解决问题,没有特定的项目需要完成。

(3) 开放用户是提供开放数据的用户^[36]。Kaggle 和 SODA 中的开放用户通常是一些政府和企业机构,LinkedUp 中的开放用户是互联网上留下开放数据的教育产品用户。上海图书馆竞赛则没有特定的开放用户,上海图书馆的开放数据来源于其馆藏资源和合作机构的资源。

(4) 开放数据平台是整合并提供开放数

据的平台。在开放数据竞赛中，参赛者需要从开放数据平台中下载或访问必需的数据。LinkedUp 的开放数据需要用户从互联网中自行获取，并未构建和提供该竞赛的开放数据平台。

(5) 参赛方是开放数据竞赛的最重要的主体，上述 4 个开放数据竞赛分别针对不同的目标参赛群体。从表 4 可以看出，Kaggle 这类开放数据竞赛平台因为有着较高的知名度和奖金，吸引了很多以排名和奖金为目标的参赛者，而上海图书馆则因为其面向数字人文的特点吸引了部分特定的参赛者。

(6) 开放数据竞赛的参赛作品大体可以分为两类：一类是解决问题的方案，包括算法和模型；另一类是具体的产品和产品原型，包括 App 和小程序等。Kaggle 侧重于寻求解决问题的方案，因此参赛作品往往是一些算法和模型。剩余 3 个案例中，SODA 的竞赛作品既包含解决问题的方案，也包含具体的产品原型；LinkedUp 和上海图书馆的开放数据竞赛则重点关注参赛者提交的产品原型，其中 LinkedUp 甚至要求竞赛作品直接投入市场，接受市场的检验。

(7) 评审方是开放数据竞赛重要的主体之一，负责对参赛作品进行评估。在 4 个案例中，除了 LinkedUp 之外评审方都决定着参赛作品的排名。

(8) 投票方是对公开参与评价的参赛作品进行投票的主体，在上述 4 个案例中，只有 LinkedUp 采取了投票的环节。

3 开放数据竞赛运作机制研究

3.1 开放数据及开放数据平台机制

开放数据是开放数据竞赛的重要部分，一个高水准和高影响力的开放数据竞赛需要由高质量的开放数据和开放数据平台作为依托。笔者从数据源、技术、流程、管理角度对上述案例进行分析，梳理开放数据和开放数据平台的运作机制。

3.1.1 开放数据特征

开放数据竞赛中开放数据形式源于全球政府和企业主导的数据开放运动，但和运动倡导

的开放数据形式存在部分差别。一般的开放数据目的多是为了提高数据的透明度以及提升个体与个体之间的协作程度^[37]。在开放数据竞赛中，开放数据使用的目的是次要目标，运用开放数据得到的解决方案或产品开放创意成了主要目标。笔者在 Krumholz 等^[38]和 D'Aquin 等^[1]的研究基础上，结合开放数据竞赛案例的特点对开放数据竞赛中的开放数据特征进行总结，结果如表 5 所示。

表 5 开放数据竞赛中的开放数据特征

开放数据的特征	说 明
可获取性和可访问性	数据必须可以通过互联网或通过接口从数据库下载，并且还必须以方便且可修改的形式提供数据
重用性	必须以允许的重用方式（包括与其他数据集混合）提供数据
重分配性	数据必须赋予用户重新分配的权利
可读性	数据必须以机器能读取的形式提供
普遍参与性	数据必须允许每个人获取、访问、重用和分配，不能针对部分人群和领域提出限制

从表 5 中可以看出，为了满足利用开放数据进行模型训练和算法改善的目的，开放数据必须具备一定的重用性和重分配性。为了满足利用开放数据开发产品原型的目的，开放数据还具备具有普遍参与性，从而有利于产品用户对于数据的访问和修改。

3.1.2 开放数据平台研究

现有的开放数据竞赛大多需要开放数据平台进行开放数据的收集、整合、存储、提供。因此开放数据平台需要在保障开放数据的高质量的同时，提供给参赛用户更好的使用体验。本文基于童楠楠^[39]对开放数据质量控制的研究和朱红灿等^[40]对开放数据平台中用户体验提升研究的基础上，结合案例分析中开放数据平台的特点，提出了开放数据平台的机制，包括数据推广、数据监管、数据共享、技术应用、用户体验和评估反馈机制，如表 6 所示。

3.2 开放数据竞赛框架及运作流程

基于对开放数据以及开放数据平台的研究，结合第二部分的案例分析，笔者对开放数

表 6 开放数据平台机制分析

机 制	说 明
数据推广	对高质量和高价值的开放数据进行推广，吸引更多的参赛者对这部分开放数据的使用
数据监管	对现有开放数据的监督，包括对开放数据形式、标准、流程和反馈监督，使平台的开放数据能够满足开放数据的特征、相关法律法规和参与用户的需求
数据共享	对开放数据资源进行整合，包括加强数据统筹工作，建设不同的企业机构、不同行业 and 不同层面的开放数据共享体系；确定不同机构的开放数据开放共享程度和开放优先级；依托大数据时代的云计算技术和云计算平台实现跨行业、跨领域的数据共享
技术应用	提高开放数据的获取和使用技术，具体包括优化开放数据的采集和优化技术，以及提高数据预处理和分析技术
用户体验	开放数据平台的用户体验包括美学体验、功能体验、效能体验和交互体验。提升用户体验可以提高参赛用户对平台的参与意愿。其中，美学体验指 UI 设计和数据可视化等功能带来的用户体验；功能体验指通过提供数据搜索、获取和调用等功能提高用户体验；效能体验指提高数据可用性、重用性带来用户体验；交互体验指通过提供不同用户之间的交互功能改善用户体验
评估反馈	对开放数据平台的现有机制进行评估和反馈，主要包括自身的评价体系构建和基于开放用户和参赛者的意见反馈，从而保障现有机制的有效运行并且不断提高参赛者的使用体验

据竞赛的框架及运作流程进行了总结，如图 1 所示。其中主办方、参赛方和评审方是开放数据竞赛最重要的主体，项目提供方和投票方是

现有的框架中较为次要的组成部分，在大部分的开放数据竞赛中并未设置。下文将基于开放数据竞赛的框架图（见图 1）对运作流程进行

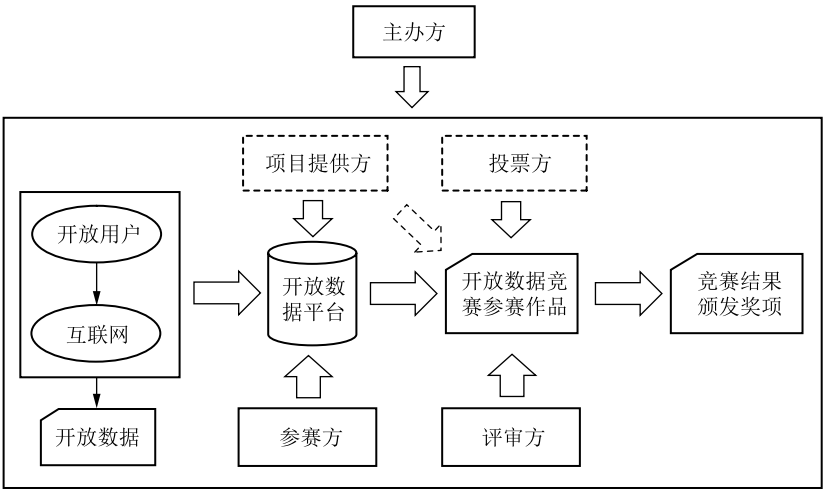


图 1 开放数据竞赛运作框架

详细阐述。

开放数据竞赛活动主要包括如下几个流程：

（1）开放数据的获取、处理与存储：一部分开放数据通过开放用户直接提供给开放数据平台，另一部分开放数据是开放数据平台通过互联网获取到的开放数据用户的日常行为数据。开放数据平台获取到开放数据后，会对开放数据进行处理和存储。

（2）项目提供方发布任务：在基于众包模式的开放数据竞赛中，存在项目提供方提供基于开放数据的任务或项目，这些任务会被提交到开放数据平台中。并非所有开放数据竞赛均包含这一流程。

（3）参赛方报名并获取开放数据：参赛方了解到开放数据竞赛的信息，并从开放数据平台中获取需要的开放数据。在基于众包模式的开放数据竞赛中，参赛方还将从开放数据平台

中选取任务或项目。

(4) 参赛方完成作品并提交作品: 参赛方基于获取的开放数据, 开发产品或完成解决方案并提交最终的作品。

(5) 评估作品: 评审方对开放数据竞赛作品进行评估。针对算法和模型类的参赛作品, 评估的标准通常是基于准确率、效率等评价指标; 针对产品及其原型等参赛作品, 评估的标准往往是一些系统评估框架, 以及投入市场使用的效果和效益等指标。

(6) 对作品进行投票: 部分开放数据竞赛设置投票方, 负责辅助评分方评价作品。

(7) 公布获奖名单并颁奖: 当参赛方提交的参赛作品经过评审方和投票方的评价后, 主办方就会公布竞赛结果并颁发奖项。

4 研究启示

开放数据竞赛是数据开放运动趋势下开放创新的新形式, 在国内外已有一些探索性的实践, 但至今尚未有较为成熟的框架模式。本文在现有的对开放数据竞赛的理论和实践探索基础上, 结合以往对开放创新机制的研究, 选取了4个国内外不同类型的开放数据竞赛案例, 进行对比分析, 并在此基础上探索了开放数据及开放数据平台的特征和开放数据竞赛的基本框架。研究结果对我国未来开放数据竞赛的筹办和推广有着重要的参考价值。

从现有的开放数据竞赛参与者角度来看, 开放数据竞赛的面向人群主要分为3类: (1) 职业参赛者, 他们参与开放数据竞赛的主要目的是追求奖金, 重点关注开放数据竞赛奖金的额度; (2) 数据科学从业者或者求学的学生, 希望能够通过参与开放数据竞赛锻炼自身能力、提高业内名声或者获取创业积累等, 他们重点关注开放数据竞赛的影响力; (3) 数据爱好者, 他们对平台提供的某类开放数据具有浓厚的兴趣, 并希望基于这些开放数据完成一些产品, 他们主要关注开放数据的质量。其中前两类是开放数据竞赛的主要参赛人群。目前开放数据竞赛的目标参与群体依然是面向开放数据专业人士, 大部分参与者对开放数据竞赛有一定的功利性目的。然而在现阶段, 国

内仅有少部分的开放数据竞赛具有较高的知名度和行业认可度, 无法满足广大数据科学相关从业者的需求。另外, 这类开放数据竞赛在开放数据平台的构建、开放数据形式以及赛程的设置方面都存在较多不足之处。除了面向专业人士之外, 开放数据竞赛也可以降低技术门槛, 针对特定领域提供部分开放数据用以吸引更多参赛者参与到创新的过程中。例如上海图书馆的开放数据竞赛通过开放大量的馆藏资源, 允许参赛者提交多种形式的开发原型, 可以吸引到很多非数据科学专业的人士参与到数字人文的价值共创过程中。基于此, 国内的开放数据竞赛可以采取上述两类发展方向: 一类是大力增加、完善和推广专业型开放数据竞赛; 另一类增加面向特定领域开放数据的非专业型开放数据竞赛, 从而更好地利用开放数据, 实现开放创新和社会进步。

从开放数据的角度看, 目前国内的开放数据平台构建尚有许多改进之处。在数据源层面, 现有的开放数据平台的数据来源多是政府和企业机构, 部分从互联网收集的开放数据也存在较多诸如合法性、格式化、完整性、可读性等问题。虽然开放数据竞赛旨在利用互联网大数据时代丰富的数据资源, 然而目前的开放数据在边界上依然需要进一步明晰。譬如, 现有的开放数据竞赛所使用的主要是有条件的开放数据, LinkedUp 开放数据竞赛通过允许开放用户自行收集开放数据来实现数据的进一步开放获取, 然而由于缺少开放数据平台对数据进行统一管护, 反而无法提供很好的开放数据服务。从技术层面看, 开放数据的分析和存储技术良莠不齐, 数据接口标准难以统一, 导致不同机构之间的数据共享门槛较高。例如在上海图书馆的开放数据平台中, 部分数据无法依靠接口调用, 技术层面还不够完善。从流程层面看, 不同开放数据竞赛对数据的生命周期的重视程度存在差异。只有保障数据的收集、存储、预处理、分析、共享等完整生命周期, 才能提高开放数据的质量。从管理层面来看, 不同案例对开放数据的边界不统一, 导致不同机构的数据开放程度不一致, 从而影响

开放数据开发者在使用过程中的效率。因此在未来的探索中需要从多方面关注开放数据的质量,鼓励更多的开放用户参与到数据开放的过程中。

从开放数据平台机制角度而言,现有的开放数据竞赛大都有开放数据平台,这些平台承担着数据推广、数据监管、数据共享、技术应用、用户体验和评估反馈的机制。然而现有的开放数据平台的上述机制都不够完善,存在着各类问题。譬如,数据推广方式较为单调、缺乏必要的数据监管、缺乏明确的开放数据共享体系和云计算技术、开放数据的技术应用手段较为落后、用户体验不够完善以及无法对开放数据平台提供评估反馈等。开放数据平台作为开放数据竞赛的技术支持平台,对竞赛的成功举办、推广和影响起着至关重要的作用。建设功能完善的开放数据平台,有利于提升开放数据竞赛的影响力和用户参与度。

从开放数据竞赛的作品角度而言,现有的开放数据竞赛正在不断探索新型的作品形式,在早期的算法和模型的基础上,增加了多种形式的产品开放,再到产品的运营。未来可以考虑增加更多形式的作品提交方式,鼓励更多行业和领域的

参赛者参与到开放数据的使用和管护中。

从开放数据竞赛运作角度而言,现有案例在运作流程上存在较大差异,竞赛主办方在筹办开放数据竞赛过程中需要花费较长时间和经济成本对运作流程进行探索,影响了开放数据竞赛的筹办效率。本文总结了必要的竞赛运作流程,为主办方提供了可行的指导建议。值得注意的是,部分必要流程的缺失会对竞赛质量造成负面影响。现有案例中,仅有部分有明确的开放数据获取、处理和存储流程。虽然开放数据竞赛重在对数据使用方案的创新,但开放数据的质量(输入)对作品质量(输出)也会产生较大的影响,主办方应当给予充分的重视。此外,对作品进行投票也是开放数据竞赛不可缺少的流程:一方面公开投票是让作品的评审流程变得更加公正透明,从而提高竞赛在业界的认可度;另一方面,公开投票加强了主办方对竞赛的宣传力度,有利于提升竞赛在业界的宣传和影响力。综合来看,开放数据竞赛应当在必要的运作流程基础上,不断扩展和完善开放数据竞赛的规则和机制,从而更好地实现开放数据的使用和价值共创。

参考文献

- [1] D'AQUIN M, DIETZE S, DRACHSLER H, et al. Building the Open Elements of an Open Data Competition[J]. D-Lib Magazine, 2014, 20(5/6): 1-11.
- [2] MORGAN L, FINNEGAN P. Deciding on Open Innovation: An Exploration of How Firms Create and Capture Value with Open Source Software[J]. Ifip-the International Federation for Information Processing, 2008, 287(89): 229-246.
- [3] PRAHALAD C L. Co-creation experience: the next practice in value creation[J]. Journal of Interactive Marketing, 2004, 18(3): 5-14.
- [4] HJALMARSSON A, RUDMARK D. Designing digital innovation contests[C]//International Conference on Design Science Research in Information Systems. Springer, Berlin, Heidelberg, 2012: 9-27.
- [5] HJALMARSSON A, HJALMARSSON A, JUELL-SKIELSE E, et al. Contests as innovation intermediaries in open data markets[J]. Information Polity, 2014, 19(3, 4): 247-262.
- [6] LEIMEISTER J M, HUBER M, BRETSCHNEIDER U, et al. Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition[J]. Journal of Management Information Systems, 2009, 26(1): 197-224.
- [7] BOUDREAU K J, LAKHANI K R. Using the crowd as an innovation partner[J]. Harvard Business Review, 2013, 91(4): 60.
- [8] SCHÖTTNER A. Fixed-prize tournaments versus first-price auctions in innovation contests[J]. Economic Theory, 2008, 35(1): 57-71.
- [9] BRUNSWICKER S, JENSEN B, SONG Z, et al. Transparency as design choice of open data contests[J]. Journal of the Association for Information Science & Technology, 2018, 69(10): 1205-1222.
- [10] BULLINGER A C, NEYER A K, RASS M, et al. Community-based innovation contests: Where competition meets cooperation[J]. Creativity & Innovation Management, 2010, 19(3): 290-303.
- [11] HUANG W. Handbook of Human Centric Visualization[M]. Berlin: Springer, 2014.
- [12] FÜLLER J, HUTTER K, HAUTZ J, et al. User

- Roles and Contributions in Innovation-Contest Communities[J]. *Journal of Management Information Systems*, 2014, 31(1): 273-308.
- [13] HANSEN E G, BULLINGER A C, REICHWALD R. Sustainability Innovation Contests: Evaluating Contributions with an Eco Impact-Innovativeness Typology[J]. *Social Science Electronic Publishing*, 2011, 5(2/3): 221-245.
- [14] WANG L, LI J, LI Q, et al. The Value Co-Creating Model and the Coordination Mechanism Research of Science and Technology Service Industry Innovation Ecosystem[J]. *Science & Technology Progress & Policy*, 2017, 34(6): 69-74.
- [15] FRANKE N, PILLER F. Value creation by toolkits for user innovation and design: The case of the watch market[J]. *Journal of product innovation management*, 2004, 21(6): 401-415.
- [16] 上海市经济和信息化委员会. SODA上海开放数据创新应用大赛[EB/OL]. [2018-10-07]. <http://shanghai.sodachallenges.com/>.
- [17] 张磊, 夏翠娟. 面向数字人文的图书馆开放数据服务研究——以上海图书馆开放数据应用开发竞赛为例[J]. *图书馆杂志*, 2018, 37(3): 33-38.
- [18] ZHAO Y, ZHU Q. Exploring the motivation of participants in Crowdsourcing contest[C]// *Proceedings of the 33th International Conference on Information Systems. AIS*, 2012b.
- [19] ZHENG H, LI D, HOU W. Task Design, Motivation, and Participation in Crowdsourcing Contests[J]. *International Journal of Electronic Commerce*, 2011, 15(4): 57-88.
- [20] HOWE J. The rise of crowdsourcing[J]. *Wired magazine*, 2006, 14(6): 1-4.
- [21] PAN M E C . The Emergence of Open-Source Software in China[J]. *International Review of Research in Open & Distance Learning*, 2007, 8(1): 1-18.
- [22] ROBERTS J A, HANN I H, SLAUGHTER S A. Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects[J]. *Management Science*, 2006, 52(7): 984-999.
- [23] OKOLI C. Information product creation through open source encyclopedias[C]// *Computing, Engineering and Information*, 2009. *International Conference on. IEEE*, 2009: 349-352.
- [24] LEUF B, CUNNINGHAM W. The Wiki way: quick collaboration on the Web[M]. Addison-Wesley Longman Publishing Co., Inc., 2001.
- [25] MAJCHRAK A, WAGNER C, YATES D. Corporate wiki users: results of a survey[C]// *Proceedings of the 2006 international symposium on Wikis. ACM*, 2006: 99-104.
- [26] PARKER K R, CHAO J T. Wiki as a Teaching Tool[J]. *Interdisciplinary Journal of Knowledge & Learning Objects*, 2007, 3(1): 57-72.
- [27] BRUNSWICKER S, JOHNSON J. From Governmental Open Data Toward Governmental Open Innovation (GOI): A Global Perspective[J]. *The Handbook of Global Science, Technology, and Innovation*, 2015: 504-524.
- [28] HENDLER J, HOLM J, MUSIALEK C, et al. US government linked open data: semantic. data. gov[J]. *IEEE Intelligent Systems*, 2012, 27(3): 25-31.
- [29] 贺文. Kaggle: 寻找数据科学界的老虎伍兹[J]. *IT 经理世界*, 2013(17): 64-65.
- [30] SRAMANAMITRA. Thought Leaders in Big Data: Interview with Jeremy Howard, President and Chief Scientist of Kaggle (Part 1) | Sramana Mitra[EB/OL]. [2018-12-18]. <https://www.sramanamitra.com/2013/03/09/thought-leaders-in-big-data-interview-with-jeremy-howard-president-and-chief-scientist-of-kaggle-part-1/>.
- [31] 上海图书馆. 开放数据应用开发竞赛2018 | 上海图书馆[EB/OL]. [2018-12-18]. <http://opendata.library.sh.cn/>.
- [32] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[M]// *The semantic web. Springer, Berlin, Heidelberg*, 2007: 722-735.
- [33] KALAMPOKIS E, TAMBOURIS E, TARABANIS K. Linked open government data analytics[C]// *International Conference on Electronic Government. Springer, Berlin, Heidelberg*, 2013: 99-110.
- [34] 夏义堃. 国际组织开放政府数据评估方法的比较与分析[J]. *图书情报工作*, 2015(19): 75-83.
- [35] 郑磊, 高丰. 中国开放政府数据平台研究: 框架、现状与建议[J]. *电子政务*, 2015(7): 8-16.
- [36] TZOVARAS B G, ARVAI K, DULANEY M, et al. Open Humans: A platform for participant-centered research and personal data exploration[J]. *bioRxiv*, 2018: 469189.
- [37] 姚乐, 樊振佳, 赖茂生. 政府开放数据与智慧城市建设的战略整合初探[J]. *图书情报工作*, 2013(13): 12-17.
- [38] KRUMHOLZ H M, WALDSTREICHER J. The Yale Open Data Access (YODA) project—a mechanism for data sharing[J]. *New England Journal of Medicine*, 2016, 375(5): 403-405.
- [39] 童楠楠. 我国政府开放数据的质量控制机制研究[J]. *情报杂志*, 2019, 38(1): 135-141.
- [40] 朱红灿, 胡新, 李顺利. 基于Kano模型的政府数据开放平台用户体验要素分类研究[J]. *现代情报*, 2018(12): 13-21.

(下转第156页)

参考文献

- [1] 戴志骞. 图书馆学术讲稿[J]. 教育丛刊, 1923, 3(6): 19-67.
- [2] 查修. 编制中文书籍目录的几个方法(民国十一年七月至十二年六月的经验)[J]. 东方杂志, 1923, 20(22): 96-110.
- [3] 馆讯 中文编目事务改进近况[J]. 清华周刊·图书馆增刊第六十四号, 1930-05-19.
- [4] 施锐. 奋斗一生——纪念施廷镛先生[M]. 南京: 南京大学出版社, 2008: 156.
- [5] 蒋元卿. 中国图书分类之沿革[M]. 上海: 中华书局, 1937: 207, 200-204.
- [6] 馆讯 修改分类法[J]. 清华周刊·图书馆增刊第五号, 1929-10-07.
- [7] 馆讯 修改分类法金石类[J]. 清华周刊·图书馆增刊第十号, 1929-10-21.
- [8] 查修. 中文书籍编目问题[J]. 新教育, 1924, 9(1-2): 215-231.
- [9] 查修. 中文书籍分类法商榷[J]. 清华学报, 1925, 2(1): 353-405.
- [10] Melvil Dewey. Decimal classification and relative index for arranging cataloging and indexing[M]. Boston library bureau, 1885: 50-227.
- [11] 朱自清. 图书馆概况[J]. 清华周刊, 1936, 向导专号卷: 37.
- [12] 不详. 清华大学图书馆工作细则[M]. 油印本, 1955.
- [13] 查修. 编制中文书籍目录的几个方法(续)[J]. 东方杂志, 1923, 20(23): 86-103.
- [14] 韦庆媛, 邓景康. 清华大学图书馆百年图史[M]. 北京: 清华大学出版社, 2013.
- [15] 中文新书目录[J]. 清华周刊·图书馆增刊第八号, 1929-10-16.
- [16] 金敏甫. 中国现代图书馆概况[M]. 广州: 广州图书馆协会, 1929: 38.
- [17] 国家图书馆《中国图书馆分类法》编辑委员会. 《中国图书馆分类法》第五版使用手册[M]. 北京: 国家图书馆出版社, 2012.
- [18] 图书馆工作报告[J]. 清华校友通讯, 1940, 6(2-3).

董琳 女, 清华大学图书馆特藏部, 馆员。研究方向: 特藏建设及民国文献分类组织。作者贡献: 提出研究思路, 撰写和修改论文定稿。E-mail: donglin@mail.tsinghua.edu.cn 北京 100084

刘春美 女, 清华大学图书馆编目部, 副研究馆员。研究方向: 文献分类编目及阅读推广。作者贡献: 提供分类编目技术理论指导和修改论文。北京 100084

(收稿日期: 2018-11-15 修回日期: 2019-06-14)

(上接第123页)

赵宇翔 南京理工大学经济管理学院, 教授, 博士生导师。研究方向: 互联网群体协作、数字人文、用户信息行为。作者贡献: 研究思路设计、部分论文撰写、论文修订。E-mail: yxzhao@vip.163.com 江苏南京 210094

全冲 南京理工大学经济管理学院, 硕士研究生。研究方向: 用户信息行为、数字人文。作者贡献: 案例分析、论文撰写。江苏南京 210094

张轩慧 南京大学信息管理学院, 博士研究生。研究

方向: 数字人文、网络信息资源管理。作者贡献: 案例调研、论文修改。江苏南京 210023

张磊 上海图书馆(上海科学技术情报研究所), 高级工程师(正高级)。研究方向: 数字人文、信息服务。作者贡献: 案例指导、论文修改。上海 200040

(收稿日期: 2019-02-03 修回日期: 2020-01-07)

(上接第137页)

等連絡会・実務者協議会. デジタルアーカイブの構築・共有・活用ガイドライン. 2017[EB/OL]. [2019-10-25]. https://www.kantei.go.jp/jp/singi/titeki2/digitalarchive_kyougikai/.

- [12] 奈良文化財研究所, 東京大学史料編纂所. 木簡・くずし字解読システム[EB/OL]. [2019-10-25]. <https://mojizo.nabunken.go.jp/>.

- [13] みんなで翻刻[EB/OL]. [2019-10-25]. [http://](http://www.honkoku.org/)

www.honkoku.org/.

三浦宽二 日本滋贺县爱荘町立爱知川图书馆, 副馆长, 日本东北大学教育学修士。研究方向: 图书馆经营论、地方文献资料论。E-mail: kodanuki1128@gmail.com 日本滋贺县

(收稿日期: 2019-11-10)