

公共数据开放的产出与效果研究

——以上海开放数据创新应用大赛为例*

郑磊** 吕文增

复旦大学国际关系与公共事务学院数字与移动治理实验室 上海 200433

摘要: 公共数据开放之后有哪些利用者? 开放数据是如何被利用的? 带来了什么产出和效果? 以上海开放数据创新应用大赛为例, 对开放数据的利用者、利用程度、受益者、利用产出和效果等问题进行了研究。研究表明, 公共数据开放后能吸引来自不同地区不同职业的数据利用者。数据利用者们对比赛专门开放数据、政府开放数据平台上常态开放的数据都有较高等度的利用, 并将这些数据和各类自带数据进行了融合利用。基于开放数据产生的创新应用形式多样, 应用领域广泛, 潜在受益者包括普通市民、政府部门和商业机构。公共数据开放具有巨大的潜在商业效益与社会效益。

关键词: 政府数据; 公共数据; 数据开放; 开放数据; 公共服务; 大数据

DOI: 10.16582/j.cnki.dzzw.2017.09.001

一、前言

公共数据开放与利用是近年来学术界与产业界的热点命题。开放政府数据是指可被任何人自由地访问、获取、利用和分享的数据, 可创造公共价值, 推动经济增长和社会发展^[1]。然而, 数据在开放之后如何被利用, 并能产生哪些应用和效果, 目前仍是开放数据研究中的薄弱点, 而其中基于我国实践开展的研究尤为缺乏。

在开放数据之后, 为推动数据的创新应用, 国内外的政府部门陆续举办了各种开放数据创新大赛。2015年, 上海开放数据创新应用大赛(Shanghai Open Data Apps, 以下简称 SODA 大赛)以城市交通为主题开放交通类公共数据, 供数据利用者开发利用, 吸引了大量团队参赛, 产生了一批参赛作品, 为研究公共数据开放的产出和效果提供了契机。

二、文献综述

公共数据开放后的产出、效果和价值是开放数据领域一个重要的研究问题。一些研究对开放数据价值产生的过程和路径开展了研究。欧盟委员会对开放数据的价值链进行了描述: 数据开放后, 将被分析和处理, 提供有价值的信息或被可视化呈现, 推动新的数据产品或数据服务的形成, 然后在此基础上诞生更具聚合性、综合性的产品或服务。在理想状态下, 公共部门数据开放将培植出多样化的产品和服务^[2]。英国开放数据研究院分析了开放数据的价值链, 认为其分为从数据发布者到数据整合者、数据利用者再到最终产品用户等四个层次, 每个层次的数据都产生了经济价值, 并提供了多种价值创造的商业模式^[3]。PWC 的最新研究以英国的开放数据挑战赛(Open Data Challenge Series)为例, 构建了开放数据的影响路径^[4]。开放数据应用大赛生成的商业创

* 基金项目: 2014 年国家自然科学基金项目“大数据背景下开放政府数据的因素与机理研究: 系统动力学建模与政策仿真”(编号: 71473048)。

** 通讯作者 收稿日期: 2017-08-31

意或商业方案,能吸引到不同程度的投资,这些都属于直接产出(outputs);而大赛的效果(outcomes)可以分为短期、中期和长期成果。短期成果包括最终的商业企划案、财务收入增加、开发产品等;中期成果包括持续的商业业务、示范效应等;长期的成果包括给经济、社会、环境带来的远期影响。世界银行《开放数据助力经济增长》的报告^[5]认为,虽然数据所创造的直接经济价值非常显著,但绝大多数开放数据所带来的效益都是间接的,也就是最大受益人是服务的终端消费者而不是服务提供者。体现在数据终端消费者身上的价值来源于数据及其结果的结合,即创意、新技术、现有服务的改进等,该报告指出评估消费者端的间接效益比评估服务提供者端的直接效益难得多。夏义堃(2016)的研究^[6]认为,开放数据的价值通过汇集、转换、分析应用开发和平台建设等一连串链条激发,每个阶段都有价值需求。郑磊(2015)^[7]认为,在创造数据价值的过程中,政府、数据利用者、普通公众和外部环境的作用缺一不可,共同决定着开放数据的最终效果,构成一个生态系统;在价值创造机理上,政府数据从开放、利用到创造价值是一个动态循环的过程。

学者们还对数据利用后产生的价值类型进行了分析。Stefan Huber^[8]认为,政府开放数据带来的主要是公共价值,公共价值就是把更广泛的公众利益带到开放数据的价值链中。Benington^[9]认为,开放数据的公共价值涉及到国家、市场与公民社会,具体包括:①经济价值。通过经济行为对公共领域产生价值。②社会与文化价值。通过提高社会凝聚力与社会福利产生价值。③政治价值。通过激发与支持民主对话和促进公共参与产生价值。④生态价值。通过降低污染与全球变暖、促进可持续发展产生价值。PWC的研究^[4]也发现,开放数据可以给经济、社会、环境带来远期影响,包括:经济方面,如

增加就业,促进总附加值的增加等;社会方面,如犯罪率降低,促进教育发展,更好地利用政府数据,促进政府透明等;环境方面,如温室气体的减排等。Michel和Niels^[10]根据价值创造过程中外部参与以及经济价值比例的高低,将开放数据的价值分为政府透明度(参与度低、经济价值比例少)、效率与效果的提升(参与度低、经济价值比例高)、公众参与和合作(参与度高、经济价值比例低)、创造新的服务与商业契机(参与度高、经济价值比例高)等四类,并认为政府应当更多地采用众包行为来帮助自身提高技术能力,通过公众教育提升公务员与公众对于开放数据价值的意识。

国内外的现有研究都发现,开放数据能在多个维度上产生价值,这些价值体现在数据开放、利用到产生应用的整个价值链上的各个阶段,研究者需要将数据利用者、数据利用程度、数据利用的产出效果和受益人等都纳入开放数据价值创造的过程中进行研究^[11]。目前,我国开展的有关数据开放的效果与价值的研究多着眼于数据开放之后的中长期、宏观价值,而对于政府数据被开放后带来的实际产出和效果尚缺乏足够的研究。本文将基于我国实际案例对该问题进行研究。

三、研究内容与方法

本文基于2015上海SODA大赛产生的作品与相关数据,对开放公共数据开放的产出和效果进行研究,包括数据利用者、数据利用程度、数据利用的产出和效果等多个维度(参见表1)。

2015年的SODA大赛由上海市经济和信息化委员会和交通委主办,以“城市交通”为主题,开放了城市道路交通指数、地铁运行数据、一卡通乘客刷卡数据、浦东公交车实时数据、强生出租车行车数据、空气质量状况、气象数据、道路事故数据、高架匝道数据、新浪

表1 研究框架

分析框架	内容	描述
数据利用者	来源	参赛团队和人员来源地
	职业	参赛人员的职业背景
数据利用程度	数据利用个数	作品中利用开放数据集的个数
	数据融合利用度	相关数据集之间的融合利用程度
数据利用产出	利用形式	参赛作品的应用形式
	利用领域	参赛作品的应用领域
数据利用潜在效果	受益人类型	参赛作品指向和服务的潜在受益人群体
	利用效益	参赛作品的潜在效益,分为商业效益和社会效益

微博数据等10类数据集,以可机读格式开放,数据总量共约1TB。数据时间周期为一个月、半年到一年不等,地域范围除少数限于个别区县外,囊括上海所有地区。主办方意在通过开放公共数据为上海城市交通征集创新应用和解决方案。该次大赛于2015年8月开放网上平台报名,报名参与人数共计2035人,初赛阶段共收到505份作品,其中100个作品进入了复赛,最终有15个作品进入决赛。研究对大赛参赛团队提交的505份作品以及参赛人员基本信息进行了定量分析。

四、数据利用者

(一) 参赛团队地区来源

研究发现,该次大赛最终共有823个团队参赛,中国大陆的团队占绝对多数(787支队伍,约占总数的95.63%)。在这787支来自中国大陆地区的参赛队伍中,来自上海的团队共429支,来自北京的团队132支,来自江苏、广东、浙江的团队都达到20支以上,参赛团队在27个省(自治区、直辖市)均有分布。总体上,北上广和沿海城市参与度最高。除中国大陆外,约4%的团队来自其他国家和地区,其中美国23支、新加坡4支、英国2支、日本1支。

1. 参赛者中国大陆地区分布

该次大赛报名参与人数共计2035人,其中1919人

来自中国大陆地区,116人来自世界其他地区。

报名参赛的中国大陆人员的来源情况,上海以1106人居首,其次为北京286人,而江苏、广东也达到100人以上。总体上,北上广和沿海城市的人士参与度较高,中西部地区也有一定程度的参与。

2. 参赛者全球分布

除中国的参赛者外,还有来自美国、新加坡、英国、澳大利亚、日本、荷兰、芬兰、加拿大、西班牙等国家和地区的参赛者。

(二) 参赛者职业分布

研究分析了参赛人员的职业(参见图1),从职业



图1 参赛人员职业分布



图4 参赛者自带数据利用程度

讯点评类平台或网站所提供的饮食、旅游、文娱活动等数据，也成为参赛者使用程度较高的数据。这些数据的利用有助于为公众提供路线规划、生活服务、商户选址等创意服务，也有利于政府进行城市交通管理。此外，停车场动态数据、道路路网、相关企业名录和地址、商户住宅数据等也成为智能交通类方案的重要数据源。

(四) 数据融合利用度

研究还评估了数据集融合利用中的相关性，数据之间的相关融合利用能够拓展单个数据集利用时的限制范围，有利于提高数据分析的深度洞察力。某一数据集和其他数据集被共同使用的次数越多，数据集之间的相关

性就越紧密。分析表明，在 0.01 置信水平上，相关系数最高的是道路事故数据和气象数据 (0.4726)、地铁运行数据与强生出租车数据 (0.4445)、实时空气质量数据与新浪微博数据 (0.4056)。

各数据集之间的相关关系如图 5 和图 6 所示。在图 5 中，和公交卡数据相关系数最高的数据依次是空气质量数据、新浪微博交通数据和道路事故数据等；和地铁运行数据相关系数最高的依次是出租车出行数据、道路交通指数和新浪微博数据；和地面道路交通指数相关性最高的三类数据依次是地铁运行数据、公交车实时数据和出租车出行数据。其他数据集间相关性如图 6 所示。

此外，研究还分析了每个数据集和与之存在显著相关性 (显著水平 < 0.01) 的其他数据集的个数。如图 7 所示，公交车实时数据集和其他九类大赛开放数据集之间都存在显著的使用相关性，而出租车出行数据集在使用相关性上并没有和其他数据集产生更多的显著关系，更多情况下是被单独使用。

研究还对参赛团队所使用的三类数据 (大赛专门开放的数据、政府数据服务网上开放的数据和参赛者自带

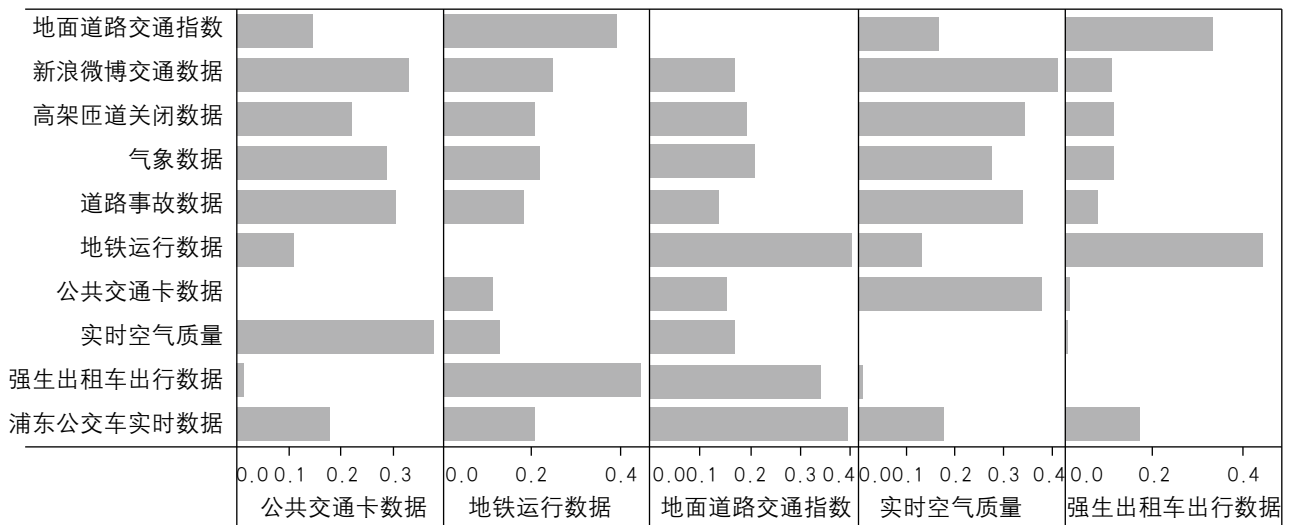


图5 数据集利用相关性一

郑磊 吕文增·公共数据开放的产出与效果研究——以上海开放数据创新应用大赛为例

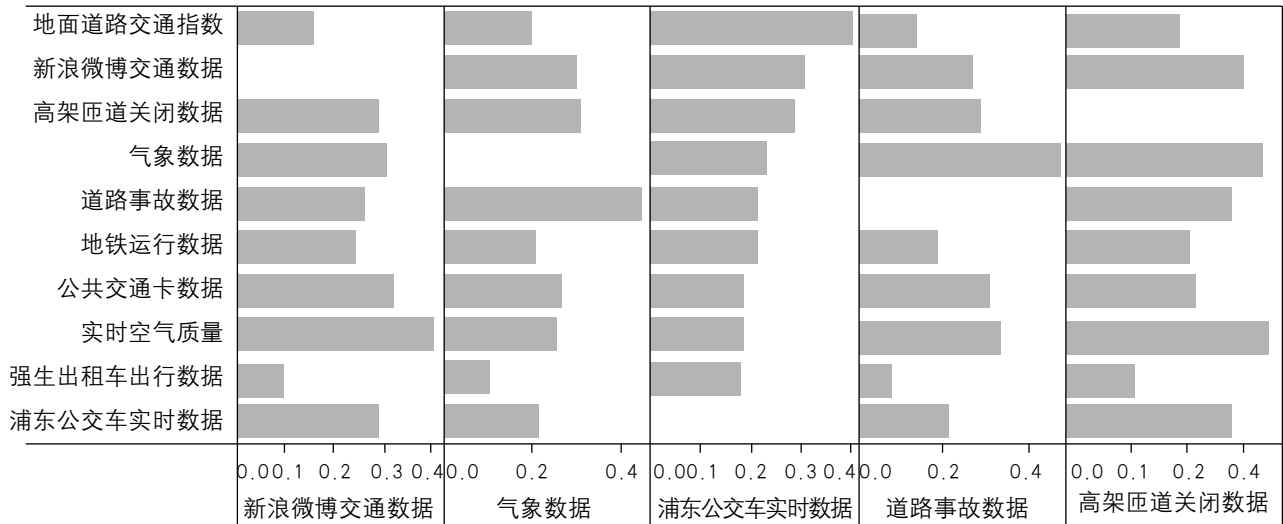


图6 数据集利用相关性二

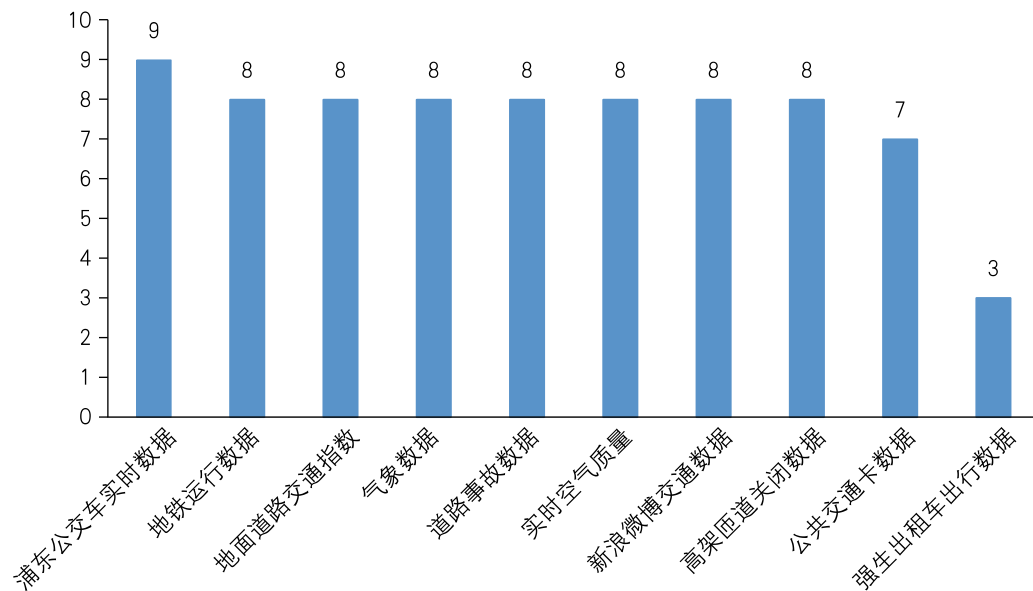


图7 数据集融合利用个数

数据) 在数据利用过程中是否进行过融合利用进行了分析。在 505 个参赛作品中, 约 77% 的创意方案将大赛十个开放数据中的一项或几项数据集进行了融合利用; 约 44% 的创意方案将大赛开放的数据与自带数据进行关联, 在参赛者自带数据的补充下, 拓宽了创意思路,

产生了多样化的应用; 约 31% 的创意方案在开放数据之间以及开放数据和自带数据之间都产生了两两关联应用; 约 25% 的方案只使用了一种数据, 没有进行任何数据融合利用, 相比之下, 这类方案的数据挖掘程度也较低, 应用领域比较单一。

六、数据利用的产出

(一) 利用形式

研究分析了参赛作品的创新应用形式。常见的应用形式是 APP、网页或技术应用方案。在 505 个案例中,有 55.9% 的团队选择了将创新应用展现在手机应用客户端 APP 上,14% 的方案选择通过网页形式展现,30.9% 的方案通过技术应用方案展现。此外,还有 7 个方案使用了微信公众号的应用形式,其他约 1.8% 的未明确说明应用形式,这类作品大多停留在创意层面,尚未落实到最终的应用层面。

(二) 利用领域

大赛产生了一批数据创新应用,服务于城市交通和其他领域,既有对现有应用的整合与优化,也有全方位的创新。研究基于初赛提交的方案,提炼出 10 种主要的应用领域,分别是导航、规划时间/路线、打车/拼车、停车、人流/拥堵的预测与规避、交通事故规避/风险预测、天气/空气质量服务、规划线路或路网/站点选址、班次调整派遣、应急管理/方案制定等(参见图 8)。

如图 8 所示,多数方案的创意描述会涉及到多个应用领域,被涉及最多的领域是为公交出行者提供出行方式、时间与线路的预测或规划,约占 40%;其次是人流与拥堵的预测和规避、路网/路线/站点选址与规划等。即使是同一类型的应用领域,产生的创意也丰富多样。如规划线路和选址涉及的创意内容包括诸如基于人流数据设计站点与路线、公共自行车站点布置、地铁站间短驳公交车规划、定制公交的线路设计、企业商家选址等跨领域与行业的多种应用方向。除主流创意外,作品中也不乏小众化、个性化的创意,如对桥梁维护保养、同城物流、基于周边交通状况对写字楼定价、上班迟到保险、基于天气情况的外贸进出口物流调整、驾驶员或车辆评级等服务。

七、数据利用的潜在效果

(一) 受益人类型

受益人是指参赛作品直接指向和服务的潜在受益群体。根据群体性质的不同,分为个人、公共管理者和商业机构。在 505 个方案中,有 71.5% 的方案将个人(普

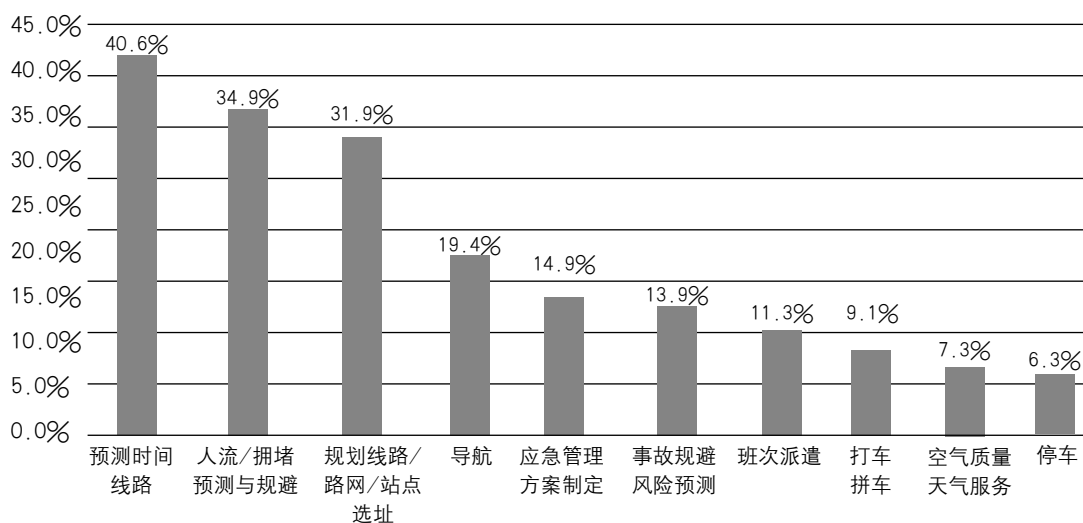


图 8 主要应用领域

富有创意和个性,其潜在受益者包括普通市民、政府部门和商业机构。研究显示,公共数据开放确实具有巨大的潜在商业价值与社会效益。

研究结果还表明,数据开放将有助于吸引和集聚数据利用人才、提升城市服务水平和管理能力、激发创新活力、促进产业发展。将公共数据作为一种公共产品开放给社会进行创新应用,发动公众参与和协作,共同解决公共问题,具有无穷潜力,可创造巨大的公共价值。需要特别强调的是,政府部门自身也是数据开放的重要受益者,政府部门需充分认识到这一点,不断提高开放数据的意愿和积极性,推动更多更好的数据向社会开放。

参考文献:

- [1] 郑磊. 开放政府数据研究:概念辨析、关键因素及其互动关系[J]. 中国行政管理, 2015(11): 4.
- [2] European Commission. Elements of A Data Value Chain Strategy[EB/OL]. (2013-07-11)[2017-08-31]. <https://ec.europa.eu/digital-agenda/en/news/elements-data-value-chain-strategy>.
- [3] Permission Granted: The Economic Value of Data Assets under Alternative Policy Regimes[R]. The Open Data Institute, 2016.
- [4] Nesta and the Open Data Institute. Open Data Challenge Series Final Report[EB/OL]. 2015[2017-08-31]. <https://www.pwc.co.uk/assets/pdf/nesta-and-the-open-data-institute-pwc-report-october-2015.pdf>.
- [5] Stott A. Open Data for Economic Growth[R]. World Bank, 2010.
- [6] 夏义堃. 开放数据开发利用的产业特征与价值链分析[J]. 电子政务, 2016(10): 41-50.
- [7] 郑磊. 开放政府数据的价值创造机理:生态系统的视角[J]. 电子政务, 2015(7): 2-7.
- [8] Huber S. The Fitness of OGD for the Creation of Public

Value[C]. Conference for Democracy & Open Government, 2014.

- [9] Benington J. From Private Choice to Public Value[EB/OL]. (2005-09-30)[2017-08-31]. <http://www.publicnet.co.uk/features/2005/09/30/from-private-choice-to-public-value/>.
- [10] Avital M, Andersen N B. The Value of Open Government Data: A Strategic Analysis Framework[R]. Orlando, U.S.: Pre-ICIS e-Government Workshop, 2012.
- [11] 徐慧娜, 郑磊, Pardo T. 国外政府数据开放研究综述:公共管理的视角[J]. 电子政务, 2013(6): 2-7.

作者简介:

郑磊, 博士, 复旦大学国际关系与公共事务学院副教授, 数字与移动治理实验室主任, 担任 Government Information Quarterly 编委, 联合国全球电子政府评估专家组成员。研究领域包括数字治理、政府数据开放、跨边界信息共享、政府社会化媒体应用、移动互联时代的公共服务等。

吕文增, 复旦大学数字与移动治理实验室研究员, 复旦大学国际关系与公共事务学院行政管理硕士。研究方向包括数据治理、开放数据、政府社会化媒体应用、移动互联时代的公共服务等。