政府大数据治理的挑战及对策

范灵俊[1]，洪学海[1]，黄晁[2]，华岗[3]，李国杰[1]
1. 中国科学院计算技术研究所信息技术战略研究中心，北京 100190；
2. 宁波中国科学院信息技术应用研究院，浙江 宁波 315048；
3. 宁波市智慧城市规划标准发展研究院，浙江 宁波 315048

摘要
大数据是城市智慧的“来源”，利用好大数据可以有效缓解或解决城市发展中的诸多问题。政府部门掌握的大数据关系国计民生，与公众生活息息相关，是高价值密度的数据，如何治理和利用政府大数据，是智慧城市建设的核心问题之一。以宁波市政府大数据为例，阐述了政府大数据治理面临的挑战，包括政府数据内部共享的需求和障碍、政府数据对外开放和利用的问题等，同时给出了相应的对策及建议。

关键词
政府大数据；治理机制；内部共享；对外开放

中图分类号：C931.2　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016028

Challenge and countermeasure of governing government big data

FAN Lingjun[1], HONG Xuehai[1], HUANG Chao[2], HUA Gang[3], LI Guojie[1]
1. Information Technology Strategy Research Center, ICT, CAS, Beijing 100190, China
2. Ningbo Institute of Information Technology Application, CAS, Ningbo 315048, China
3. Ningbo Academy of Smart City Development, Ningbo 315048, China

Abstract
Big data is the source of the city's wisdom, making good use of big data could solve or relieve the problems caused by the city's development. The big data managed by the government are of high value because they are closely related to people's life and public service. How to govern the government big data and explore its value is one of the key challenges during the construction of smart city. Taking Ningbo as a case study, the challenges of governing government big data were described, which include the needs and obstacle of sharing big data by internal government, and the problems of opening government big data and helping to bring it about market utilization, solutions and policies were also proposed.

Key words
government big data, governing policy, internal sharing, opening

1 引言

智慧城市的建设和健康发展离不开数据的采集、整合、分析和利用，城市“智慧化”建立在数据的充分开发利用的基础之上。政府大数据是智慧城市建设中一笔宝贵的“资产”。随着信息技术的发展和政府部门信息化工作的推进，近年来，政府部门积累了越来越多的数据。与互联网大数据和行业大数据相对应，这些数据被称为“政府大数据”。众所周知，政府大数据关系到国计民生，与百姓的生活密切相关，被认为是高价值密度的数据，利用好政府大数据，对政府决策、经济发展和公共服务水平的提升等都大有裨益。政府大数据在智慧城市建设中具有不可替代的关键作用。

政府各部门是政府大数据的实际掌握者，是政府大数据价值链的“源头”。由于管理体制的限制以及政府部门信息化系统建设的历史原因，政府大数据分散在各个部门，甚至各个系统当中，被称为“信息烟囱”或“数据孤岛”。这种碎片化的数据现状不利于政府大数据的开发利用，因为大数据的奇妙就在于，不同来源、不同类型的数据融合之后，可能挖掘出原来发现不了的结果或结论。因此，政府大数据的融合是政府大数据治理的首要问题。然而，政府的各个部门到底掌握了多少数据、数据的类型有几种、哪些是“死”数据、哪些是“活”数据、动态更新的频率如何、数据质量如何等，都是政府各个部门需要首先搞清楚的问题，即摸清自己的“数据家底”，也是实现政府大数据融合的前提。

政府大数据融合的目的是对其进行利用，发挥政府大数据的价值。政府大数据的利用分为内部利用和外部利用。所谓内部利用，是指政府部门内部之间对政府大数据的利用，而外部利用是指政府大数据对外开放，取得社会化利用。内部利用需要解决好政府大数据的共享问题，外部利用需要解决好政府大数据的对外开放问题。开放数据后，众多第三方创新企业可以对开放的原始数据进行加工和传播利用，创造前所未有的新价值，形成过去没有的数据生态链。例如，空间地理信息可用于采矿、林业、农业、渔业、能源、航海、交通运输等行业，气象信息可用于农业、旅游业、灾难管理、环境评估等业务。美国纽约市政府数据公开后的2年时间内，就有500多家企业做数据相关服务。美国于2000年取消了对民用GPS精度的限制，国内就有约300万个就业岗位依赖于GPS。

综上所述，从政府大数据的“源头”——政府部门出发，政府大数据治理的流程包括：摸清各个部门的数据家底、梳理各个部门的数据需求、实现各个部门数据的融合、建立政府部门内部的数据共享机制、建立政府大数据对外开放的模型，最终实现政府大数据的市场化利用，释放政府大数据蕴含的巨大价值。

2 政府大数据治理的内涵及意义

政府大数据是指政府部门掌握的大数据，如人力资源和社会保障部门掌握的社保数据、卫生部门掌握的医疗数据、交通部门掌握的交通数据等。与此对应的有企业大数据——企业掌握的大数据，如阿里巴巴公司掌握的消费者数据、腾讯公司掌握的用户数据等，这是根据数据掌握者的不同来进行划分的。从这个意义上讲，政府大数据只是众多大数据中的一种，特指政府部门多年来，在信息搜集、信息系统

建设及业务办理过程中累积的大数据。本文讨论的“政府大数据治理”的对象是政府大数据。对全社会各种大数据的治理同样需要政府制定战略、出台政策、制定法规、实施项目等，这是更高层面的问题，不在本文探讨的范围内。

研究表明，政府掌握的数据占到国家掌握的数据的70%~80%，如果将这部分数据利用起来，数据开发者能利用政府大数据开发创新性应用，提供更好的服务，创造更多价值，能推动经济增长乃至整个经济增长方式的转型，提高整个国家在大数据时代的竞争力。

政府是提供公共管理和公共服务的部门，政府大数据与民生和经济都密切相关，利用好政府大数据，不仅可以提高政府的决策水平，如公共交通的配备和供给；而且可以服务于社会经济发展，如修建商场的最优选址；同时还可以提升政府公共服务能力，如突发事件的预测、预防等。由此可见，政府大数据是政府部门的一笔重要的资产，不能让这笔资产“沉睡”。盘活政府大数据资产，发挥政府大数据的价值是一项重大的系统工程，需要全方位、多层次、多角度的共同发力。政府部门作为数据的掌握者，处在政府大数据价值链的源头，起着决定作用。本文从政府大数据治理的源头出发，探讨政府大数据治理的内涵、遇到的问题及对策。

政府大数据治理的根本目的是发挥政府大数据的价值。除了加强政府部门内部对政府大数据的利用，还要将政府大数据对外开放，取得政府大数据的市场化利用，从而形成政府大数据的产业链和价值链。数据是互联网创新的重要基础，政府大数据提供给社会进行增值开发和创新应用，可以激发大众创业、万众创新。

普遍认为，大数据具有4V（volume、variety、velocity、value）特征，其中之一便是大数据的类型多样、来源广泛。由于政府部门的条块分割及业务类型的不同，政府大数据恰好具备了这个特点。要发挥政府大数据的价值，首先要对政府部门的数据进行整合，实现不同部门数据的融合。

多个部门的数据整合并不是最终目的，整合的目的首先是实现政府内部数据的互联互通和共享，满足最优的政府职能的行使，依靠大数据分析提高管理能力和决策水平；其次是实现政府大数据的对外开放，供给社会进行市场化利用。这两者都有一个共同的问题是找到核心应用，满足政府管理和公众的实际应用需求。

目前比较普遍的现象是，即使公开的信息也没有及时地推送到公众的眼前，即信息是公开了，但没有形成稳定的信息推送服务系统，只在网站上“一挂了之”，没有实现“最后一公里”。目前政府部门的数据开放平台（网站）大多是从政府部门自身出发给出的。其列出的可开放数据清单只是政府数据的一部分，甚至是很小的一部分。大多数公众对一些公开的数据并不感兴趣，或者根本就不知道政府数据开放平台的存在。然而并不能就此认为公众对政府数据漠不关心。因为数据清单是政府从自身出发列出的，在政府看来可开放的，而不是全部数据。公众可能对“单采血浆站许可证审核信息”不感兴趣，但不表示他对“医院医护人员专家信息”不感兴趣。

此外，政府大数据的开放和利用不能理解成将原始数据裸露给公众。举例来说，对于培训机构的基本信息，市民兴趣可能不大，但如果将培训机构做成一个大众点评模式，公众能看到附近有哪些培训机构、各个机构的优势在哪里、大家对它的评价如何等，那样会大大增加公众的兴趣。因此，政府大数据开放和利用不能将原始数据裸露给市民，而必须加以应用，找到应用

场景。裸露原始数据给公众，既有安全和隐私问题，公众也不一定感兴趣。对政府数据应用而言，企业是政府和公众的桥梁，特别是“微、小、中”企业。这些企业将政府大数据汇集的资源与公众的各种各样的需求进行对接，并开发出各类应用，这比政府将大数据直接裸露给公众更有效。

3 国内外政府大数据治理实践与研究现状

政府大数据作为战略资产已经得到越来越多国家的重视。截至2014年，全球已有63个国家和地区推动政府大数据的开放，并开放了超过700 000个政府数据集。通过开放政府数据促进社会转型，带动大数据产业的发展，已经成为各国的普遍共识[1]①。

以美国为例，通过数据开放，美国2013年在政府管理、医疗服务、零售业、制造业、位置服务、社交网络、电子商务7个重点领域产生的直接和间接价值已经达到了2万亿美元。截至2014年2月10日，美国数据门户网站开放了88 137个数据集、349个应用程序、140个移动应用；美国新版的数据开放门户网站将原来的金融、企业和安全等六大类数据集拓展至农业、消费、教育、能源等20大类，与经济和民生需求相关的数据集大幅增加。

英国政府通过高效地使用公共大数据的技术，一方面优化政府部门的日常运行和刺激公共机构的生产，另一方面在福利系统中减少诈骗行为和错误的数量，包括更有力度地追收逃税漏税的税款②，从而节省了大量开支。以医疗领域为例，2009 年英国公开数据研究所在英国医疗领域发起了开放数据运动，将全国医生所开的处方信息连接在一起，放在网上分享。病人在接受治疗时可以自主判断采用不同的处方，从而一年为英国减少了2亿英镑的药费浪费。

西班牙政府为了发展大数据、建设智慧城市，把古老的港口城市桑坦德作为试点③。目前，桑坦德市中心已经安装了近1万个传感器，覆盖面积约为6 km^2。有些传感器安装在路灯、电线杆和建筑物墙壁上，隐藏在灰色小盒子里，还有的传感器甚至被埋在停车场的沥青地皮下。这些传感器每隔几分钟就会把城市的交通、天气、行人动作等数据传输到坎塔布里亚大学的实验室——整个城市的数据收集中心。市民只要在智能手机上下载安装“城市脉搏”程序，就可以获得整个城市的相关详细信息。此外，传感器还将具备帮助公园优化花草灌溉用水量的功能，避免水资源浪费。传感器会事先通知清洁工哪个垃圾箱需要清空，他们就不必每天沿着街道查看每个垃圾箱是否需要清空了。

同时，西班牙政府将推动更多政府信息公开化，公众可以获取很多以前比较机密或难以得到的信息，包括人口结构变化和房地产价格变动的统计数据。这些海量数据可以激励程序员们创造出更多的应用程序，让桑坦德变得更加智能化。

与西班牙类似，韩国也首先从大数据基础设施上发力④，韩国政府宣布将建设一个开放大数据中心，该中心面向中小型企业、风险企业、大学和普通公民，他们都可以通过该中心对大数据进行提炼和分析，利用大数据技术解决业务或者研究方面的问题。

同时在对外开放和利用政府数据方面，韩国政府努力打造“首尔开放数据广场”。“首尔开放数据广场”目前包含33个数据库、880个数据集，为用户提供十大类的公共数据信息，包括育儿服务、公共交通路线、巴士到站时间、停车位、各地区天

① http://opendatachina.com/the-business-case-for-open-data/

② http://intl.ce.cn/specials/zxgjzh/201207/10/t20120710_23475902.shtml

③ http://www.china-cloud.com/dashujuzhongguo/disanqi/2014/0116/22745.html

④ http://www.china-cloud.com/dashujuzhongguo/disanqi/2014/0110/22652.html

气预报及餐厅推荐等涵盖生活方方面面的信息，并配有图表、数据、地图和网络链接等。首尔市还计划根据开放应用程序接口的利用情况和数据的再利用情况，向市民和企业进一步开放有用数据。

总结起来，国外政府大数据的治理主要关注3个方面。第一是重视和完善基础设施的建设，包括布局和设置各类传感器、提高宽带和网络水平、建设大数据中心、开放平台等；第二是推动和加大政府大数据的开放，包括公共机构的各种数据集，通过共享平台、访问接口等方式呈现给企业和个人，以机器可读方式优先发布高价值数据；第三是吸引和鼓励企业和公众对政府大数据进行开发利用。开放政府大数据的最终目的是取得市场化利用，释放政府大数据的价值。因此采取激励措施鼓励企业和创新者利用开放数据开发应用，发展数据产业是发达国家的普遍做法。

国内学者对政府大数据的治理及共享也做了大量的研究。有参考文献探讨了我国政府大数据共享的3种模式：日常共享、重要共享和专项共享⑤。孙艳艳等[2]分析了我国开放政府数据当前取得的成果和存在的问题，从顶层设计、政府责任主体、信息化建设、民间参与、意识形态等5个方面对我国开放政府数据的发展策略进行了探讨。郑磊等[3]通过样本对目前国内各地的政府数据开放融合共享实践进行了评估⑥，指出了我国政府数据开放实践当前存在的6个方面的主要问题：数据量少、单一价值低、可机读比例低，开放的多为静态数据，数据授权协议条款含糊，缺乏便捷的数据获取渠道，缺乏高质量的数据应用，缺乏便捷、及时、有效、公开的互动交流。蒋定福等[4]总结了我国政府信息资源共享发展历程，并从政府信息资源管理机制、安全存储机制、监督机制、评价考核机制、法律保障机制5个方面论述了大数据的政府信息资源共享保障机制。李宇[5]分析了当前制约信息整合与共享的主要因素，包括技术标准和业务标准各异、体制不顺与部门利益分割等。陈真勇等[6]提出了一种新的智慧城市数据共享和融合框架——智慧城市数据互联框（smart city linked data framework），该框架将大数据处理分为数据存储层、数据转换层、数据互联层和数据共享层，可广泛应用于交通数据、地理信息数据、天气数据等。于鹏[7]针对传统对非结构化数据缺乏有效利用的问题，提出了基于数据仓库与大数据融合的解决方案。李卫东[8]将政府信息资源共享分为4个层次：政府各职能部门之间的资源共享、不同层级政府之间的资源共享、国家权力机关与政府之间的资源共享、政府与企业之间和政府与公民之间的资源共享以及提出公共数据中心法、虚拟数据中心法、开放式的Web服务法3种共享方法。

4 政府大数据治理的需求及挑战

政府大数据的治理是一项系统工程，不仅涉及技术，还涉及政策、环境、法律法规、公共管理等多个方面。政府大数据治理的目标就是发挥出政府部门掌握的大数据的价值，服务经济发展、提升公共服务、提高管理水平，形成政府大数据的价值链。作为政府大数据价值链的源头，政府部门在政府大数据的开发利用上起着不可替代的作用。“源头”的问题首先要解决好，才能把政府大数据价值链的上下游打通。本节以宁波市政府大数据为例，选取10个与民生相关的部门，对其掌握的数据进行走访调查、总结和对比，重点分析和梳理了政府大数据在融合、共享、开放及

⑤ http://www.e-gov.org.cn/egov/web/article_detail.php?id=153729

⑥ http://chuansong.me/n/1388604

利用方面面临的需求和挑战。

表1对宁波市10个政府部门掌握的数据情况（包括数据的更新情况）进行了摸底和总结。各部门的数据情况截止到2015年8月，根据各部门的业务特点，**表1**中只列出了可以公开并且可以摸清的数据，应当指出，各部门实际的数据拥有量要大于**表1**中列出的数据量。

从**表1**可以看出，政府部门掌握着大量高价值密度的数据，然而尚未发挥出真正意义上的大数据价值。不同于一般互联网的数据量大、价值密度低的特征，政府大数据与公众生产、生活的方方面面都密切相关，是具有高价值密度的数据。每个政府部门都有不同的数据存量，涉及不同的价值方面和业务类型。

同时，调查发现，大多数政府部门对政府大数据的价值都有充分的认识，不同部门根据对数据应用的设想和规划对其他部门的数据需求也很强烈。以宁波市公安局和81890求助服务中心为例，**表2**和**表3**分别选取了这两个部门对其他部门的部分数据需求。

从**表2**、**表3**可以看出，公安局和81890

表1 宁波市10个政府部门掌握的数据概况

政府部门	掌握数据情况
人力资源和社会保障局	医疗保险管理信息系统（总参保人数达到近1 000万人）； 社会保险统一征缴信息系统（基本养老保险参保320多万人（大市540多万人），城乡居民养老保险参保40多万人（大市近170万人）；工伤参保170多万人（大市290多万人），生育参保160多万人（大市250多万人））
卫生局	全市各医院现有HIS（医院信息系统）数据约5 000 GB，年增长约500 GB； 现有电子病历系统数据约100 GB，年增长约20 GB； 现有LIS（实验室信息系统）数据约2 500 GB，年增长约500 GB； 现有PACS（影像归档和通信系统）数据约2 000 TB，年增长约400 TB； 疾病控制数据：现有数据容量600 GB，年均增长量约100 GB； 卫生监督数据：各类数据几千几万条不等；120急救数据现有数据容量：数据库600 GB，电话录音450 GB，无线对讲录音450 GB，年均增长量：数据库70 GB，电话录音80 GB，无线对讲录音50 GB； 血液管理数据：现有数据容量约6.5 GB，年均增长量1 GB左右； 病理数据：现有数据容量：82 GB，年均增长量：20 GB
公安局	人口基础数据库（已导入1 006.3万条实有人口信息，包括421.2万流动人口和585万户籍人口）
规划局	基础地理信息数据（数字线划图数据、数字高程模型数据、数字正射影像数据、遥感影像数据）； 地理空间框架数据库（电子地图数据、地理实体数据、地名地址数据、城市三维模型和2.5维数据、360度街景影像）
统计局	按月从10多万家调查对象收集、汇总近千个经济和社会发展指标数据，并在此基础上，通过部门数据收集、数据分类汇总及整合等途径生产并对外提供近400个主要指标及1 000多个分类指标，再加上年报数据和各阶段性的国情国力调查，每年新增数据量达1 GB，累计数据总量20 GB以上
质量技术监督局	以企业档案为基础的业务信息库（企业档案69 413份，累计日常巡查记录44 255份，监督抽查报告35 203份，检验项目达820 590项，行政许可证书信息16 609份，案件信息11 456件）
教育局	数字化教育教学资源（本中心未包含大学园区图书馆）1.4 TB，每周进行更新，每年新增600 GB； 管理类数据信息10 GB，每半年进行更新，每年新增1 GB
海曙区经济和信息化局	截至2015年3月，政务信息资源中心数据总数已达2 592万条，历史数据14 384万条，总计16 976万条； 涉及人口数据表82张，组织单位数据表224张，地址数据表14张； 数据字段：涉及人口属性字段1 069个，组织单位属性字段2 561个，涉及地址属性字段169个
81890求助服务中心	从数据格式上可以分为语音数据、图片数据、视频数据、数据库数据、网页及其他可编译脚本数据五大类； 从应用角度可以分为81890应用业务系统数据、81890录音数据、81890系统平台功能数据、81890网站数据、81890历史资料数据五大类 截至2015年3月，81890求助服务中心共产生数据记录4 400.24万条，共储存文件9 889 569份，数据总容量16.12 TB
民政局	社会救助信息管理平台、居民家庭经济状况核对系统、医疗救助即时结报系统、民间组织管理信息系统、重点优抚对象医疗一站式结算平台、死亡人员核对管理信息系统、地名信息管理服务平台等

表 2 宁波市公安局当前对其他部门的数据需求清单（部分）

提供单位	采集项目
市贸易局	商业类从业人员和会员卡信息
市工商行政管理局	企业法人及员工信息
	服务行业从业人员信息
	家政服务行业信息、从业人员信息
	维修登记行业员工信息
	企业动产抵押登记
	商标注册信息管理局
	婚姻中介
市卫生局	医院门诊信息、住院信息、健康档案信息、体检信息、医院从业人员信息、精神患者排查信息、无偿献血人员信息、儿童体检信息
	民办医疗机构企业信息、就诊人员信息
市教育局	学校学生信息、家长信息、教师信息、招生信息、学校信息
	民办教育培训机构信息、从业人员及学生信息
	校车驾驶员信息
市城市管理行政执法局	环卫工人信息
	供水信息
	燃气充装单位信息和用户信息

表 3 81890 求助服务中心对其他政府部门数据需求清单（部分）

提供单位	81890求助服务中心需求信息
市工商行政管理局	企业及个体工商户注册、注销相关信息（如联系电话、地址、经营项目），各类商会信息、广告、展览审批等业务信息
市公安局	人口户籍信息及证照验证（验证志愿者身份、核实鹊桥会未婚男女的单身身份）、出入境申办信息、交通管制信息、交通法规咨询、交通设施报修、交通路线咨询、路况信息、各派出所/治安大队/看守所/监狱信息等；另诸如车辆年检预约是否能向81890开放，让百姓享受到免费的服务
市规划局	城市地图信息、兴趣点信息（81890求助服务中心也能根据百姓求助的内容标注新的兴趣点）
市卫生局	各大医院、卫生院、社区服务站、诊所及妇幼保健、民办医疗机构等基本信息（包括医院联系方式、特色、先进仪器、擅长领域、专家坐诊、门诊时段），体检、工伤鉴定、防疫、医疗纠纷、药品（保健品）、药店信息、献血、急救等方面的查询、咨询、政策法规等信息；建议卫生系统的预约挂号资源能向81890求助服务中心开放，因为81890求助服务中心与妇儿医院合作，可代居民在妇儿医院挂号，不收取任何额外费用，深受百姓的欢迎，而通过其他平台进行的网上挂号均需要付出费用
市人力资源和社会保障局	涉及人力资源和社会保障方面的政策、查询、考证等信息：如退休档案、社保、补贴等查询，人事代理、补助发放、优惠政策企业托管、职业培训、招聘、职称技能资质考证、验证等以及相关政策法规的查询
市教育局	学校/幼儿园的基础信息、教学特色、招生地段划分及要求、出国留学、培训机构及其培训内容，托管机构、书籍查购、教育讲座、自考、成考、教学仪器销购、对外开放场馆等相关信息
市住房和建设局	拆迁安置、保障房、房产交易、住房贷款、物业、房屋鉴定评估、房屋安全（如白蚁防治）、老小区公共部位维修、产权归属、违章搭建、新楼盘、房屋档案、地段划分规划、建筑垃圾清理等方面的相关政策、数据、办理指向、房产中介、建筑企业、相关专业培训机及培训项目等信息
市民政局	养老机构、福利院、救助站、所属彩票网点、社团组织、补助发放、困难户等信息
市城市管理行政执法局	水、电、气、路灯、户外霓虹灯各网点信息、业务办理、保修抢修、停用预告、公园广场活动等信息
市电业局	各网点信息、业务办理、保修抢修、停用预告、基站辐射

求助服务中心对政府大数据的应用都有很好的设想，对其他部门的数据有巨大且比较明确的需求。由此可见，仅从政府部门内部的需求来看，加大对政府大数据的治理，推动政府大数据的融合、共享和利用势在必行。

然而，我国政府大数据的共享和利用，当前最大的挑战在于缺乏政府数据共享的统一标准和规范，缺乏治理机制设计。比如，大多数政府部门，只要是掌握了一部分核心数据的，都希望把其他一些部门的数据整合过来，形成一个大一统的数据资源平台，对开发各种应用都有一个蓝图和规划。存在的问题是，各部门都不愿意把自己的核心数据交出去。因为有些数据就是某些部门的命脉，一旦交出去，这个部门可能就没有存在的必要了。因此一些政府部门将政府大数据和信息资源产权部门化，设置信息利用的壁垒。这也是“数据孤岛”、“信息烟囱”存在的主要原因之一。

5 政府大数据治理的对策及建议

5.1 政府大数据治理流程与框架

政府大数据治理是一项系统工程，需要多方的共同努力。政府部门作为数据的实际掌握者，是政府大数据价值链和产业链的源头，在政府大数据开发利用过程中起着决定作用。**图1**从政府部门出发，提出了政府大数据的治理流程。首先各个部门要摸清本部门的数据家底，有多少数据存量、是什么类型的数据、更新频率如何、是否涉密等，这是实现政府大数据融合的前提。然后建立统一的数据资源平台，实现政府大数据的融合。数据融合之后，一方面可以实现部门内部共享，另一方面，可以对外开放，实现市场化利用，从而发挥政府大数据的价值。

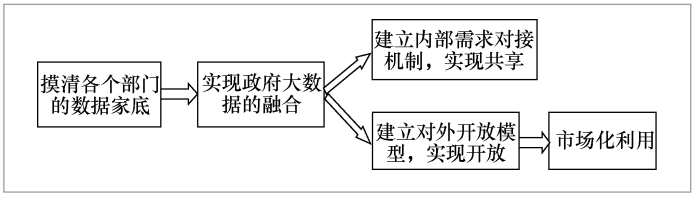

图1 政府大数据治理流程

对于大数据而言，每种数据来源都有一定的局限性。只有融合集成各方面的原始数据，才能反映事物的全貌。事物的本质和规律隐藏在原始数据的相互关联之中。不同的数据角度不同，但可以描述同一件事情。政府的数据可能更客观、更全面，网络上的数据可能更反映民意。对同一个问题，不同的数据可以反映互补信息，增加了解的广度和深度。要实现各方面信息的集成，首先要做的就是数据的共享和融合。如图2所示，数据融合是实现部门之间数据共享的前提。

政府各部门除了摸清和建立自己的数据清单，还要结合自己的业务需要及大数据应用需求，提出对其他部门的数据需求清单，有了以上两份清单，政府可以成立更顶层的大数据应用协调部门，进行数据的对接和协调，甚至是数据的交易。解决挖掘核心应用、确定主导部门、避免重复建设等问题。如**图3**所示。

需要成立顶层协调部门统一协调本市的大数据发展。通过出台统一的大数据建设标准体系，包括数据标准、接口标准，规范各单位大数据项目的开展，便于市、各单位数据资源共享、应用、挖掘、关联、分析和决策，为各单位在大数据项目规划、设计、建设过程中提供专业指导。

各个部门还要对数据进行分级分类，哪些数据可以开放、哪些数据不能开放、哪些数据可以逐步开放、哪些数据绝对不能开放等。因为业务性质的不同，不同的部门需要

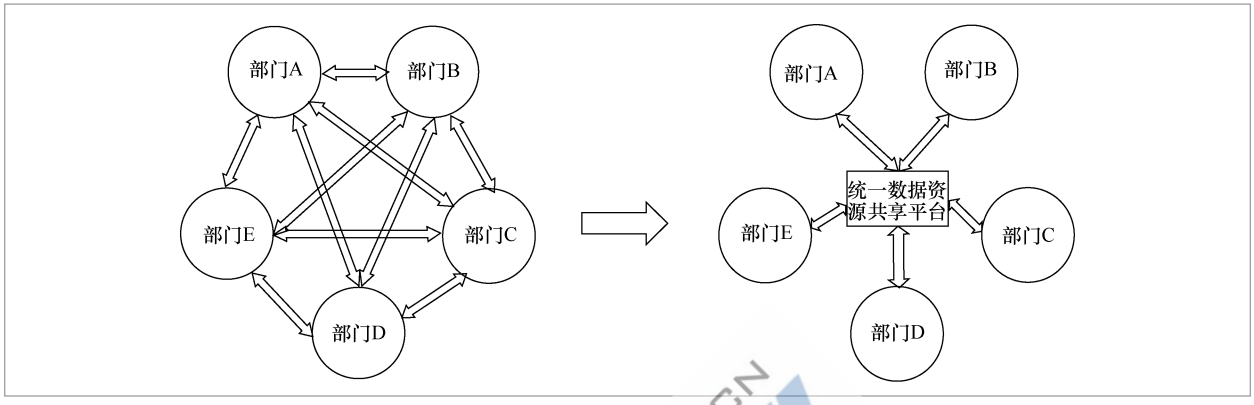

图 2 建立统一数据资源平台

不同的划分标准和依据。比如统计局，目前已经有自己的数据开放模型和标准，哪些数据可开放、哪些数据整理完才能开放、以什么样的形式公开、多长时间公开一次，都有依据。统计局有依据，其他掌握了数据的核心部门也可以有依据，并且应该有依据。

为了制定全国统一的政府数据开放标准和规范，中央有关部门（如中央网络安全和信息化领导小组办公室）应尽快组织有关专家和管理干部提出指导性意见。对数据的分级管理，政府各部门不能坐等开放标准和依据的出台。在国家层面的数据分级管理标准出台之前，各部门可以先行先试，不断总结经验。如图4所示，各部门可以根据现有的法律法规和依据，对本部门掌握的数据进行分级分类，从而建立起政府大数据分级逐步开放的模型，然后在实践中不断完善。

企业是对政府大数据进行市场化利用的主体，完成政府大数据开发利用的“最后一公里”。政府首先要对企业开放数据使用的行政许可，包括原始数据的使用；其次要保证数据的稳定性，包括数据更新的及时性和正确性。在数据应用模式上，政府建大数据的融合平台和基本库，然后向企业开放接口，企业可以根据实际需求设计算法和模型，并通过接口的方式对数据进行处理和利用。

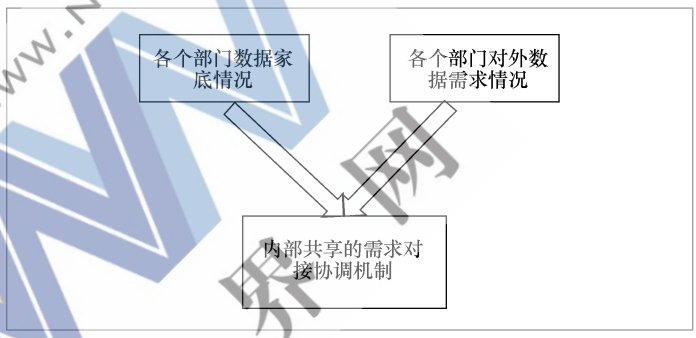

图 3 建立内部共享的需求对接机制

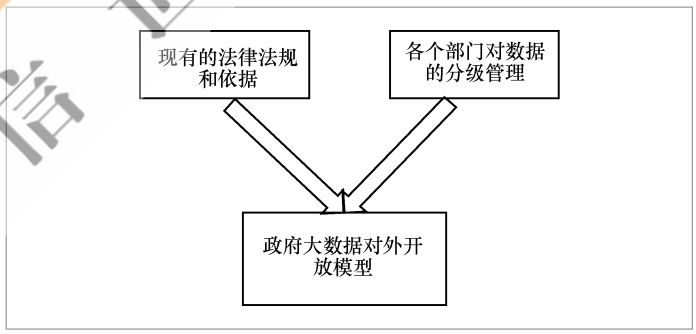

图 4 建立政府大数据分级逐步开放模型

5.2 技术要点

5.2.1 数据融合技术

如前文所述，对于政府大数据而言，实现政府大数据的融合是内部共享和对外开放利用的前提。大数据的本质特征是，整体集成的价值大于各部分价值的总和。数据融合是对各种异构数据提供统一的

表示、存储和管理，以实现逻辑或物理上有机地集中。也就是要以一种统一的数据模式描述各数据源中的数据，屏蔽它们的平台、数据结构等异构性，实现数据的无缝集成。常用的数据融合方法包括数据转换方法（联邦数据库系统）、数据聚合方法（中间件模式）、析取/转换和装载方法（数据仓库模式，简称ETL）。

在大数据兴起之前，数据仓库（data warehouse）是流行的数据融合技术。数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。只有建立统一的数据仓库，才便于对不同地区、不同部门的治理能力进行横向比较以及沿时间轴进行纵向深入研究，让连续检测、分析、计划、决策成为可能，使得数据的潜在价值得到最大限度发挥。

大数据兴起以后，深度机器学习（deep learning）成为对融合的数据进行分析的热门技术。通过机器学习、统计分析、社交网络分析、图像/视频分析、情感与舆情分析等技术手段，对多源异构融合的海量城市数据进行过滤、提取、汇聚、挖掘和展现，并通过参考历史数据和领域知识、考虑事件间的相关性和上下文感知，对事件发生成因和发展规律进行分析推理，最终给出决策支持信息。

5.2.2 防泄密与隐私保护技术

对开放政府大数据进行市场化利用，可能挖掘出国家机密或个人隐私是一个比较普遍的问题和障碍，因此需要发展大数据环境下应对数据挖掘的防泄密与隐私保护技术。开放数据需要经过对数据的脱敏保护隐私。经常采用的方式是匿名化或去标识化。能唯一标识个人的属性称为标识符，通过组合能够以较大概率标识个人的属性称为准标识符。去标识化不一定能够做得彻底，有研究表明，只要有性别、出生年月以及邮编这3个数据（准标识符），就很有可能把个人的信息还原出来。美国麻省理工学院的研究人员发现，如果获悉一天之中个人在4个不同的时间点所处的不同地点，把这个人找出来的可能高达95%。

为了测试隐私保护技术的有效性，学术界开始研究“去匿名技术”，即通过多数据源的相互匹配实现重新标识。如同加密与解密是一对孪生技术一样，隐私保护和“去匿名”也是一对孪生技术，通过技术的博弈可增强开放数据的安全性。现在已经出现不少隐私保护技术来满足不同的隐私需求，如避免被确认某人是否在某个数据集中，或某人是否有某个特殊属性，或某记录是否对应某人。

为了提高隐私保护的水平，先后提出了k-anonymity、L-diversity、T-Closeness匿名技术。在敏感属性不够多样化时，上述方法仍然有被攻击的可能。另一种匿名化技术是差分隐私，即把噪声加入数据集中，但仍保持它的一些统计属性，使之支持典型的机器学习方法。所有隐私保护方法都是以牺牲原始数据的质量来获得高匿名性。因此，在隐私安全性和数据可用性之间要做好平衡，当噪声大到一定程度时，数据可用性会变差。从效益考虑，隐私保护的成本一定要低于数据本身的价值。

5.2.3 数据定价技术

政府大数据是一种战略资产，是一种财富。资产与财富都有定价问题。为了促进数据开放和利用，应制定合理的数据价格政策。数据和信息的一个重要区别就是，信息是具有特定意义的数据，因为特定意义的存在，所以信息价格比较好估计。但数据在还没使用的时候其价值是不确定的，就像玉石，不把它剖开谁也不知道它的价值。更重要的是，数据可以反复使用，当将其应用于不同领域时可能产生超出其采集

时预期的价值。因此，一般不为买断式的数据产权定价，而是为一次使用形成的价值定价，先使用后定价，使用次数越多的数据估值越高。

数据作为一种商品，还需要考虑其稀缺性。数据的稀缺性在某种程度上反映了数据的价值密度。当一类数据很稀缺时，高估值会激励公司和个人收集和产生类似的替代数据，从而由市场进行价格调节。数据共享和交易中的一个重要挑战是防止劣质数据。多方数据相遇时，如果一方混入了劣质数据，将影响最终结果的价值。

6 结束语

政府大数据治理是智慧城市建设的核心问题之一，政府大数据是高价值密度的数据，开放和利用好政府大数据对于政府决策、经济发展、公共服务等多个方面都大有裨益。本文以宁波市政府大数据为例，探讨了政府大数据在价值链“源头”的治理机制和路径，包括通过摸清各个政府部门的数据家底和数据需求清单建立数据内部共享机制，促进数据的融合和内部利用；在中央有关部门的指导下，根据已有法律法规对部门数据分级分类逐步建立对外开放模型，促进政府大数据开放和市场化利用。此外，数据融合技术、数据防泄密与隐私保护技术、数据定价技术等都是政府大数据进行开放和市场化利用的基础并且核心的技术，需要给予足够的重视。

参考文献:

[1] 闫建，高华丽. 发达国家大数据发展战略的启示[J]. 理论探索，2015(1)：91-94.
YAN J，GAO H L. Revelation of the development strategy of big data in developed countries[J]. Theoretical Exploration，2015(1)：91-94.

[2] 孙艳艳，吕志坚. 中国开放政府数据发展策略浅析[J]. 电子政务，2015(5)：18-24。
SUN Y Y，LV Z J. Brief analysis on the development strategy of China's open government data[J]. E-Government，2015(5)：18-24.

[3] 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.
ZHENG L，GAO F. Research on China's open government data platform：framework，current situation and suggestions[J]. E-Government，2015(7)：8-16.

[4] 蒋定福，岳焱. 基于大数据的政府信息资源共享模式探讨[J]. 合作经济与科技，2015(14)：184-185.
JIANG D F，YUE Y. Discussion on the model of government information resources sharing based on big data[J]. Co-Operative Economy & Science，2015(14)：184-185.

[5] 李宇. 网络时代政府信息资源共享瓶颈因素分析[J]. 北京行政学院学报，2014(3)：65-68.
LI Y. Analysis on the bottleneck factors of government information resources sharing in the network age[J]. Journal of Beijing Administrative College，2014(3)：65-68.

[6] 陈真勇，徐州川，李清广，等.一种新的智慧城市数据共享和融合框架-SCLDF[J]. 计算机研究与发展，2014(2)：290-301.
CHEN Z Y，XU Z C，LI Q G，et al. A novel framework of data sharing and fusion in smart city-SCLDF[J]. Journal of Computer Research and Development，2014(2)：290-301.

[7] 于鹃. 数据仓库与大数据融合的探讨[J]. 电信科学，2015，31(3)：159-163.
YU J. Discussion on integration of data warehouse and big data[J]. Telecommunications Science，2015，31(3)：159-163.

[8] 李卫东. 政府信息资源共享的原理和方法[J]. 电子政务，2008(1)：65-67.

LI W D. The principle and method of government information resources sharing[J]. Chinese Public Administration, 2008(1): 65-67.

作者简介

范灵俊(1983-),男,博士,中国科学院计算技术研究所工程师,信息技术战略研究中心战略研究主管,美国韦恩州立大学访问学者,主要从事计算机体系结构、处理器设计、信息技术发展战略、智慧城市、大数据等方面的研究工作,发表论文20余篇,合作出版专著一部,申请发明专利2项。

洪学海(1967-),男,博士,中国科学院计算技术研究所研究员,信息技术战略研究中心常务副主任,主要从事高性能计算、信息服务计算以及信息技术与信息化发展战略等方面的研究工作。发表文章40余篇,合著中文专著5本。

黄晁(1972-),男,博士,中国科学院计算技术研究所副研究员,宁波中国科学院信息技术应用研究院院长,主要研究方向为视频处理技术、智慧城市系统、大数据等。在国际国内学术会议及期刊发表论文数十篇,申请国家发明专利十余项,其中4项被国家标准采纳为标准技术。

华岗(1977-),男,博士,宁波市智慧城市规划标准发展研究院副研究员,主要研究方向为智慧城市、大数据、智能可视化管理。

李国杰(1943-),男,博士,中国工程院院士,现任中国科学院计算技术所首席科学家、信息技术战略研究中心院士研究员,曙光信息产业股份有限公司董事长,中国计算机学会名誉理事长,国家信息化专家咨询委员会信息技术与新兴产业专委会副主任,中国科学院学位委员会副主席,中国科学院大学计算机与控制学院院长,中国科学技术大学计算机科学与技术学院院长等,主要从事计算机体系结构、并行算法、人工智能、计算机网络等方面的研究工作,发表论文100多篇,合著英文专著4本,出版了报告论文集《创新求索录》。先后获得国家科学技术进步奖一等奖、二等奖,首届何梁何利基金科学与技术进步奖等奖项。

收稿日期: 2016-02-20

基金项目: 中国工程院重大咨询基金资助项目(No.2014-ZD-01);中国工程院与宁波市政府重点咨询课题基金资助项目(No.2015WT002)

Foundation Items: Major Consulting Project China Academy of Engineering (No.2014-ZD-01), China Academy of Engineering and Ningbo Municipal Government Key Consulting Project(No.2015WT002)