



文献 DOI:

10.11922/csdata.840.2017.0118

数据 DOI:

10.11922/sciencedb.372

文献分类: 社会科学

收稿日期: 2017-01-19

开放同评: 2017-02-15

发表日期: 2017-07-10

2015 年科学研究数据共享现状调查

张丽丽^{1*}, 赖茂生²

1. 中国科学院计算机网络信息中心, 北京 100190;

2. 北京大学信息管理系, 北京 100871

*通讯作者 (Email: zhll@cnic.cn)

摘要: 为探讨科学数据共享问题, 设计了面向我国科学研究数据共享现状的调查问卷, 并于 2015 年依托问卷星平台面向国内科研人员展开调查。实际回收有效样本 370 份, 可为摸清当前科学研究数据共享情况提供一手材料, 亦可归档为年代基准数据, 用于后续研究。

关键词: 科学研究数据; 科学数据共享; 调查统计

数据库 (集) 基本信息简介

数据库 (集) 中文名称	2015 年科学研究数据共享现状调查		
数据库 (集) 英文名称	Survey on the current state of scientific data sharing in mainland China (2015)		
数据作者	张丽丽、赖茂生		
通讯作者	张丽丽 (zhll@cnic.cn)		
数据时间范围	2015 年		
地理区域	中国大陆	数据量	370 条
数据格式	*.doc, *.xls		
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/372		
基金项目	国家社科基金项目“政府信息资源管理创新的理论与方法” (11BTQ032)		
数据库 (集) 组成	数据集由“科学研究数据共享现状调查”问卷样表和有效样本数据 2 部分组成。问卷样表由 12 节共计 31 个问题组成; 有效样本数据包括样本和编码 2 张表单, 其中样本 370 条, 每条样本数据共 44 个字段; 编码表单定义了 44 个字段的名称、类型、宽度、定义标签和值域。		

引言

数据开放共享是当前国际热点话题。尤其随着大数据资源的战略地位日益提升, 数据驱动的科学发现已成为全球新型科研范式。《促进大数据发展行动纲要》和《“十三五”国家信息化规划》等一系列国家政策文件明确提出强化信息资源共建共享、加快释放数据红利等内容。然而, 科研数据共享仍存在诸多难题, 共享瓶颈集中在共享文化、保障机制、共享动机和共享能力等方面^[1-2]。为此, 结合博士研究生毕业选题, 笔者面向国内科研人员投放“科学研究数据共享现状调查”问卷, 试图摸清国内当前科研环境中科研主体对开放科学数据的态度观点, 为更

好地推动科学数据开放共享提供统计数据支持。

1 数据采集和处理方法

本数据集的取得由量表开发、数据采集、数据清洗和质量检验等步骤组成。

1.1 量表开发

量表开发过程主要包括设计初稿、专家评审与落实定稿等内容。

首先，根据前期文献调研并结合科研第四范式、信息共享理论、公共物品理论、新制度经济学理论以及信息生态理论、数据权利等方面的基础理论框架与相关研究成果^[1]，本调查将“我国科学研究数据共享”焦点问题的影响要素归纳为共享文化、保障机制、共享动机和共享能力四大方面，并进一步提出二级分类测量指标，详细内容见表 1，据此形成调查问卷样稿核心内容。

表 1 科学研究数据共享现状潜在影响要素量表

序号	潜在变量	测量指标
1	共享文化	a. 具有分享科学数据的传统
		b. 社会（或科研共同体）倡导数据开放共享理念
		c. 具有科学数据共享要求（如来自上级部门或资助机构等）
2	保障机制	a. 具有足够的经费支持数据共享活动
		b. 具备科学数据共享所需的基础设施与环境条件
		c. 建立了科学数据共享实施的框架方案
		d. 建立了科学数据跨平台共享与利用的标准规范
		e. 制定了完整的科学数据共享政策法规体系
		f. 建立了恰当的科学数据共享激励和评价机制
3	共享动机	a. 乐于将所拥有的数据进行公开共享
		b. 数据共享将带来科学声誉的提高
4	共享能力	a. 具有分享科学数据的技术能力和工具
		b. 具有足够的经费支持数据共享活动
		c. 具备科学数据共享所需的基础设施与环境条件
		d. 制定了科学数据共享实施的具体方案

其次，依托潜在影响要素量表，结合定量与定性手段，将问卷主体内容设计分为三大部分。第一部分为定性量表，重点描述科学数据共享资源与环境的客观情况；第二部分为定量量表，重点描述受访科研人员对国家、组织层面科学数据共享现状与制度方面的主观感受；第三部分为受访者基本信息，就此形成问卷初稿。值得补充的是，实际返回数据样本中还包括第四部分，即样本回收相关信息，如提交答卷时间等。

此外，根据问卷初稿，定向邀请科学研究领域 5 位专家学者进行试填写，并通过他们的反馈意见对问卷中指向不明、表述不清或有遗漏的方面进行完善。

最后，综合专家意见及作者观点，形成最终定稿的问卷。

1.2 数据采集与清洗

- 1. 明确调查对象：本调查面向科学研究数据共享活动核心参与力量，即国内高校、科研机构、政府科研部门和企事业单位从事科学研究的相关人员展开。
- 2. 问卷发放：本调查集中于 2015 年 3 月初依托问卷星平台(www.sojump.com)，采取随机抽样方式展开在线问卷调查。
- 3. 问卷回收：截止 2015 年 3 月 31 日共回收问卷 415 份。为保证调查问卷质量，对回收样本的相关条件做了进一步限定：①提交答卷使用验证码，防止恶意软件干预；②设置完整性限制，只回收填写完整的问卷；③根据本调查所设置题量情况，删除全部答题时间在 90 秒以下的反馈(结合笔者自测情况、综合考量后，将作答时间设置为 90 秒)；④设置同一电脑/手机限制，防止受访者重复提交答案；⑤问卷星系统自动删除个别异常答卷（如前后问题答案选项高度相仿的答卷等）。全部筛选条件框定有效回收问卷样本为 370 份，其中委托问卷星定向回收有效问卷 210 份，采取公共平台推广回收有效样本 160 份（详见数据集“来源”字段数据）。
- 4. 数据清洗：在完成质量检验、确认问卷回收数据样本具备足够的可信度和可靠性前提下，根据实际需要，进一步对拟发布数据集进行整理。其中，为保护受访者信息，过滤掉易对号入座的 IP 地址信息，并将“IP 来自”一项的详细数据加工整理为所在省份数据值，以便于了解样本分布。合并“来源”及“来源详情”两项为“来源”，便于查看数据回收渠道分布。同时针对各项返回值进行规范化编码，形成完整规范的数据集。

1.3 样本数据人口统计特征

本调查问卷所获得的 370 份有效样本的人口统计特征如下（详见数据集“受访者基本信息”相关项）。其中，受访者年龄分布方面，30~65 岁之间的科研中坚与骨干力量占比 87.3%，从事科学研究的工作年限方面具有 10 年以上丰富经验的比例约为 37.84%，一年以下工作年限的仅为 1.08%。受访者职称分布方面，具有高级职称（副高及以上）的受访者占比 40.27%，中级职称占 47.57%，中级以下职称为 12.16%。所属研究领域分布方面，自然科学约占 1/3，应用科学超半数，人文社会科学约为 14.59%（详见表 2）。

表 2 样本人口统计基本特征描述

样本特征	具体选项	样本数	百分比
年龄	30 岁以下	46	12.43%
	31~40 岁	202	54.59%
	41~50 岁	93	25.14%
	51~65 岁	28	7.57%
	66 岁及其以上	1	0.27%

（受访者基本信息）

Q10	Q11	Q12
2	4	2

3 数据质量控制和评估

3.1 质量控制

样本回收的质量控制主要包括系统限定和人工干预两部分。其中，依托问卷星平台所进行的样本回收条件系统限定详见第 1.2 节的说明。人工干预内容主要包括邀请领域专家对量表设计予以把关，笔者人工查回收样本的完整性等。其中，“IP 来自”返回值中存在 3 个“未知”值，经查对应 IP 地址均为境外，判断为合理情况，故予以保留。

此外，调研过程中，为了能够更加真实地了解科学研究数据共享实际情况，侧重增强了面向高级以上职称受访人群的问卷投放。实际回收的 370 份有效样本中，具有高级职称(副高及以上)的受访者占到总数的 40.27%、中级职称占 47.57%、中级以下职称占 12.16%。

3.2 质量评估

（1）信度检验

本研究采用 SPSS 20 版本软件处理数据，利用针对李特式量表开发的 Cronbach’s Alpha 系数值对问卷中 39 项数值型字段进行可靠性检验（表 4）。检验结果显示该问卷整体 Cronbach’s Alpha 系数为 0.884，各项主要研究测量变量的系数均超过 0.6，其中共享文化测量变量具有可信性，共享动力、共享能力变量很可信，保障机制变量为十分可信。由此可见，本问卷具有较强的信度值，可展开进一步统计分析。

表 4 问卷的可靠性统计

变量	问题项数	样本量 (N)	Cronbach’s Alpha
共享文化	3	370	0.662
保障机制	6	370	0.945
共享动机	2	370	0.702
共享能力	4	370	0.892
问卷整体	39	370	0.884

（2）效度检验

采用 KMO（Kaiser-Meyer-Olkin，取样足够度）样本测度和 Bartlett 球形检验对变量进行相关性检验测试，各项因子有效性检验结果见表 5。科学数据共享四个潜在影响变量的 KMO 值均达到或超过 0.5，且各项检测变量的 Bartlett 球形检验选项显著性 $p<0.05$ ，这表明上述变量通过有效性检验。

表 5 科学数据共享潜在影响要素有效性检验

KMO和Bartlett的检验			KMO 和Bartlett的检验		
KMO		0.589	KMO		0.907
Bartlett 球形检验	近似卡方	198.739	Bartlett 球形检验	近似卡方	2153.384
	df	3		df	15
	Sig.	0.000		Sig.	0.000
(a) 共享文化影响要素有效性检验			(b) 保障机制影响要素有效性检验		
KMO和Bartlett 的检验			KMO和Bartlett的检验		
KMO		0.500	KMO		0.838
Bartlett 球形检验	近似卡方	128.002	Bartlett 球形检验	近似卡方	874.664
	df	1		df	6
	Sig.	0.000		Sig.	0.000
(c) 共享动机影响要素有效性检验			(d) 共享能力影响要素有效性检验		

综上，本问卷设计的各项内容均通过了信度和效度检验，表明调查数据集具有一定的可信度和可靠度，适合展开进一步分析工作。但受制于资源与条件，该抽样调查问卷回收样本的随机性不可避免的受到笔者所在人际网络的影响。调查本身仍属宏观范畴，并未面向具体学科领域展开同一问卷的横断面样本比较调查。

4 数据使用方法和建议

本数据集可为了解我国科学研究数据共享进展提供一手资料，可供数据资源共享政策与实践研究使用，亦可为后续年代的数据共享进展调查提供横向比较的基准数据。可使用 Excel 综合统计，概括当前我国科研环境及科学数据共享诸方面发展态势。也可以使用 SPSS 等软件对所关心的问题交叉分析。例如依据样本人口统计特征分层，可获取年龄、科研工作年限、职称与研究领域等指标下的数据共享情况，以供横向比较分析等。

致 谢

感谢在完善调查问卷量表开发中贡献智慧的所有专家以及调查问卷中贡献观点的受访者。

数据作者分工职责

张丽丽（1984—），女，研究生，助理研究员，研究方向为科学数据开放治理。
主要承担工作：问卷设计与数据整理。

赖茂生（1946—），男，研究生，教授，研究方向为信息资源管理、信息检索。
主要承担工作：问卷设计与数据质量控制。

参考文献

- [1] 张丽丽. 面向开放服务的科学数据共享模式研究[D]. 北京: 北京大学, 2015.
- [2] 黎建辉, 吴超, 张丽丽, 等. 科学数据出版调查与分析[J/OL]. 中国科学数据, 2016, 1(1). DOI: 10.11922/csdata.120.2015.0009.

论文引用格式

张丽丽, 赖茂生. 2015 年科学研究数据共享现状调查[J/OL]. 中国科学数据, 2017, 2(3). DOI: 10.11922/csdata.840.2017.0118.

数据引用格式

张丽丽, 赖茂生. 2015 年科学研究数据共享现状调查[DB/OL]. Science Data Bank, 2017. DOI: 10.11922/sciencedb.372.

Survey on the current state of scientific data sharing in mainland China (2015)

Zhang Lili^{1*}, Lai Maosheng²

1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, P. R. China;

2. Information Management Department, Peking University, Beijing 100871, P. R. China

*Email: zhll@cnic.cn

Abstract: This paper introduces a survey focusing on the current state of scientific data sharing in current state of mainland China. The survey had been carried out in 2015, and 370 valid samples were received. This dataset can cast light on the development of scientific data sharing in China which also provides basic archiving materials for similar researches in the future as well.

Keywords: research data; scientific data sharing; survey and statistics