北京大学经济与管理学部　通讯周报

视　野

第 32 期

学部办公室编辑　　2017 年 04 月 05 日周三

编者按

在信息爆炸的时代，开放的数据共享大大加快了科学技术向前推进的步伐。相反，如果研究者在数据方面资源匮乏、壁垒森严，则反过来阻碍了科学研究的发展。目前，研究人员和机构获得数据的渠道比较多元，有的是政府公布的公共数据，有的是自建平台收集整理的专项数据，有的是从企业或第三方购得的商业数据，无论是哪种渠道，数据对于研究人员来讲就像水之于鱼，数据直接影响成果，数据质量直接影响成果质量，数据的重要性已经是不言而喻。然而，可能也是因为数据如此重要，使得数据获得不宜、共享困难，直接导致了数据质量不高、碎片化严重、发掘程度有限，大大阻碍了科学研究的发展和创新。对于学部而言，很多教研人员关注的问题相关，研究的领域相似，对于数据的掌握和需求也相近，如果能进一步促进协同合作、资源共享，必能实现互通有无、合作共赢，不断提升数据质量和完整性，整体上提升学部研究成果水平和影响力。

数据共享是推动科学研究发展的动力[1]

在信息爆炸的时代，开放的数据共享大大加快了科学技术向前推进的步伐。相反，如果研究者在数据方面资源匮乏、壁垒森严，则反过来阻碍了科学研究的发展。

数据是从事科学研究的源泉

在这个大数据时代，生物信息数据库和数据共享都有所发展，可用的技术资源也越来越多。直到最近，这种开放科学的趋势一直在帮助中国科学家获得国际竞争力。然而，公共数据的获得有困难，数据共享的机制不完善，没有有效的数据，很多科学研究无从谈起，这些问题将拖累科学研究发展和创新。

开放公共数据库并提高数据质量可以提高政府事务的透明度，这些年也一直在努力，社会关注度也越来越高。比如环保局从 2014 年起开始全面公布空气污染数据，便是一个典型的例子。之前，环保局只公布简略的日报，如今每小时都更新数据。这些数据被中国的政府部门广泛使用。根据这些数据，卫生局得以向民众发布预警，教育局也有依据决定学校是否在空气重度污染时停课，交管局则得以调整限行规定。充分披露这些数据最重要的作用，也许在于提高了公众对不断恶化的污染问题的关注程度。

受到限制的不仅仅是中国国内产生的数据，查阅国外的学术资源有时也在技术上很难实现——一些政府信息管理部门用技术手段设限，以过滤潜在的有害信息。

可能在数据获取上的不容易，使得研究人员在认识上也存在壁垒，研究团队不愿意公开自己的数据，因为数据就是无形资产，可以使科学家在自己的学术领域中获得竞争优势。上海海事大学交通运输学院

[1] 部分摘录自：“重新分析数据，研究结论可能完全不同”，原文由狼医生发表于《果壳科学人》2014-09-10 和 “数据获取不畅通，中国科学家做研究也受影响”，原文由 Zheng Wan 发表于《果壳科学人》2015-05-04（橡胶万岁/译）。题目为编者另加，内容经编者编辑整理。
我们学部内部参考，若公开引用发表，须经果壳科学人的同意。

的 Zheng Wan 在《自然》上发表的评论中就坦言：很多人都向我自己的研究团队请求海事数据——例如我们汇编的港口统计数据和船队信息——但是我们也不情愿分享这些信息。为了将零散的数据收集、整理成可用的形式，我们付出了巨大的人力物力。如果我们不公开这些数据，就可以使用这些数据撰写论文。但如果我们能更便捷地从其他来源获取优质数据，我们就会更愿意分享自己的数据。

数据质量直接影响研究成果质量

有时公布的数据，因为数据收集存在缺陷而质量不佳。这其中最典型的例子就是关于中国国内生产总值（GDP）的争议。官方公布的全国数据，与 31 个省级行政区的 GDP 总和存在着显著的差异——而且这差异还在拉大。位于北京的国家统计局承认，各省使用的统计方法并不一致，正在进行协调。

当前，教学科研人员很难获得高质量的国内数据，数据共享意识也比较淡漠。一些政府部门强化自己对数据的垄断，使研究者想获取数据变得更加艰难。如果想收集数据，需要建立数据平台收集数据，但是每个组织收集的数据都很难与别人共享，而每个组织或者研究者能收集的数据又极其有限，可从事的研究便受到限制。人文社科研究者所受的影响尤其明显，环境科学，公共健康等领域也受到影响。

然而，对于科学研究而言需要更多领域的数据。科学家感兴趣的数据，公众不一定感兴趣。政府公布的往往更多是面向公众的民生数据，而研究团队或个人的数据更专业、更学术。研究人员的数据的获得现在有些碎片化，别人分享的数据有时经过处理，有时只是一部分，这样的数据不够缜密，数据不准确即便研究方法再科学严谨，也常常容易造成偏颇和误导。这就需要研究者之间更多的交流共享。在数据分享上，相信 1+1 一定大于 2！

数据的多重分析会产生更加精准有效的研究成果

斯坦福大学医学院的研究者发现，在对以往临床研究进行的数据重分析中，大约有 1/3 得出了与原始研究不同的结论，而这可能会直

接影响到临床决策。这项新研究于今天发表在《美国医学会期刊》（JAMA）上。

研究者称，目前大多数研究机构都不愿意共享数据，从而造成数据的二次分析非常少见。他们在过去 30 多年内所发表的文献中检索，最终只纳入了 37 项发表的二次分析研究，而其中只有 5 项是由与原作者无关的独立研究者进行的。

“研究者提供他们的原始数据给其他人进行分析，这确实是非常必要的。” 斯坦福预防研究中心的负责人，医学教授约翰•伊奥尼迪斯（John Ioannidis）提到，“如果不能获得原始数据，分析工作也就难以进行。现在，对已发表研究的‘信任危机’愈演愈烈，我们搞不清发表的研究结果是否真的可信并可以用于临床问题的决策。而近期对于奥司他韦是否有用的争论只是其中的冰山一角。”

奥司他韦即“达菲”，是市场上的一种抗病毒药物。虽然奥司他韦已经获得了用于治疗甲型流感及乙型流感的许可，但一些随后进行的分析和研究发现，它所带来的收益似乎并不足以盖过副作用的风险。

伊奥尼迪斯是这项研究的资深作者，他也是近期启动的斯坦福荟萃研究创新中心（Meta-Research Innovation Center，METRICS）的负责人之一。这个组织旨在评估及优化科学实践，并以此方法来推动科学研究更好地发展。在这方面，加强数据的再现以及数据共享会是非常有益的。

论文第一作者易卜拉欣（Ebrahim）和他的同事采用 MEDLINE 数据库来进行他们的研究。MEDLINE 是一个书目数据库，它是由美国国立医学图书馆运营的，其中包括从全球约 5600 个期刊中引用的超过 2500 万条生物医学参考文献。研究者们搜索了对此前论文数据进行二次分析的英文文献。其中不包括荟萃分析，以及与原文研究目的不同的研究。

研究者们筛选了将近 3000 篇备选论文，并阅读了其中 226 篇的全文。在这当中，38 篇论文被纳入此研究中，其中一篇包括两项二次分析。其中 2 项分析随后被剔除，因为这些论文所针对的原始研究文献无法获得。在本研究评估的 37 项二次分析中，有 32 项的论文作者与原始研究存在重叠。

研究者们发现，对于“哪些患者可以从治疗中获益”这个问题，有 13 篇二次分析数据的论文（35%）都得出了与原始研究不同的结论：其中 3 篇论文中得出的治疗适用人群不同于之前的结论，1 篇文章认为更少的患者需要得到治疗，而另外 9 篇则指出应该让更多的患者得到治疗。

原始研究与二次分析结论不同，是由于二次分析中采用了不同于之前的统计和数据处理方法。一些二次分析的研究也发现了此前原始研究中的错误，比如纳入了本应该从研究中排除的患者。

例如，一项研究是有关食管静脉扩张合并出血治疗的，原研究发现血管硬化治疗不能预防再出血，但可以减少死亡率。而二次分析采用了不同的危险因素统计模型，总结出这项治疗确实可以预防再出血，但是它并不能减少死亡率。这个新的结论提示，这项治疗手段最好用于存在再出血风险的患者当中，而不是要首先考虑那些高死亡风险的患者。

另外一项研究中比较了贫血患者的不同治疗方案：刺激红细胞生成的药物采用固定剂量每 3 周给药一次，或是根据体重计算每周的剂量。在二次分析当中，研究者采用了最新更新的标准来判断何时需要开始治疗，结果推翻了原始研究的结论。

伊奥尼迪斯提到：“在二次分析的论文中，有很高比例得出了与原文不同的结果，这可能有一部分要归结为人为因素。我的意思是，在目前的科研环境中，假设进行二次分析而又得到与原文相同的结果，那么论文会很难发表。然而，将原始数据进行共享依然十分重要，这不但使研究人员可以对原论文的假设进行检验，同时也为更多的研究和数据合并分析提供了条件。”用这种方式，已知的原始数据可以用来探索新的临床问题，甚至还可以减少人们耗费在新临床研究上的精力。

伊奥尼迪斯补充道，研究者进行二次分析时得到不同的结论，这并不一定意味着原始研究存在偏倚或是造假。相反，这一结果向我们显示了进行原始数据共享的重要性，因为这样可以促进交流和共识达成，同时减少科学研究只注重“出乎意料结果”的风气。

“我非常赞成数据共享，同时相信应该鼓励独立研究者对这些数

据进行二次分析，” 伊奥尼迪斯说，“他们可以有很多独到的见解。”

（学部办公室选编）

《视野》通讯周报主要选编和交流有关人才培养、学术研究、学科建设、跨域发展等方面的前沿信息、理念、规则、机制等新见解、新做法，以便相互激励，开阔视野，启发思路，促进工作。

看完本期《视野》，若有一些想法看法，可在学部网群中进行交流。各位老师如有符合《视野》定位的文章，也欢迎推荐给学部办公室，发送至经管学部邮箱 gjgxb@pku.edu.cn。

（注：学部网群有：学部学术委员会微信群、邮件群、经管学部部务会成员邮件群、经管学部邮箱 gjgxb@pku.edu.cn）。