中国科学院科研人员的开放数据

需求调查报告

中国科学院文献情报中心　开放资源建设团队

2016 年 4 月 6 日

一、 问卷调查背景

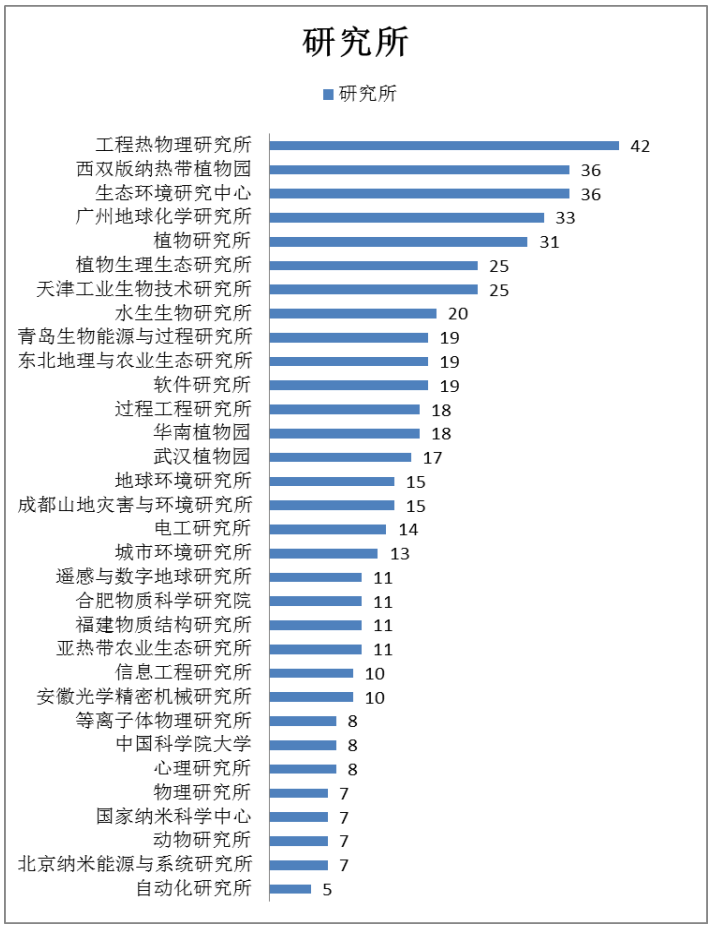
在数据密集型科研环境下，大量开放文献或网络资源的数据已被越来越多的科研人员利用，用于做研究、做文章，重新发现更多的研究成果。在学科馆员对我院科研人员的需求调研中，发现科研人员对数据的需求变得越来越强烈。很多科研人员表示，在进行科学研究、撰写论文过程中，有利用数据的好想法，却不知道到何处获取可靠的数据源；有数据需要处理，由于数据类型和数据数量太大却没有技术方面的支持，而束手无策；与 IT 公司合作，因为技术人员没有相应的学科背景而导致沟通不畅，且收费昂贵。

二、 问卷调查目标

数据定制服务，是中国科学院文献情报中心面向科研人员的数据需求，基于当前快速发展的开放信息资源，为科研人员打造的个性化数据服务，在近期正式推出。本次调查的目标在于较为全面而深入地了解科研人员在数据利用过程中的问题和需求，为该项服务的顺利开展提供依据。

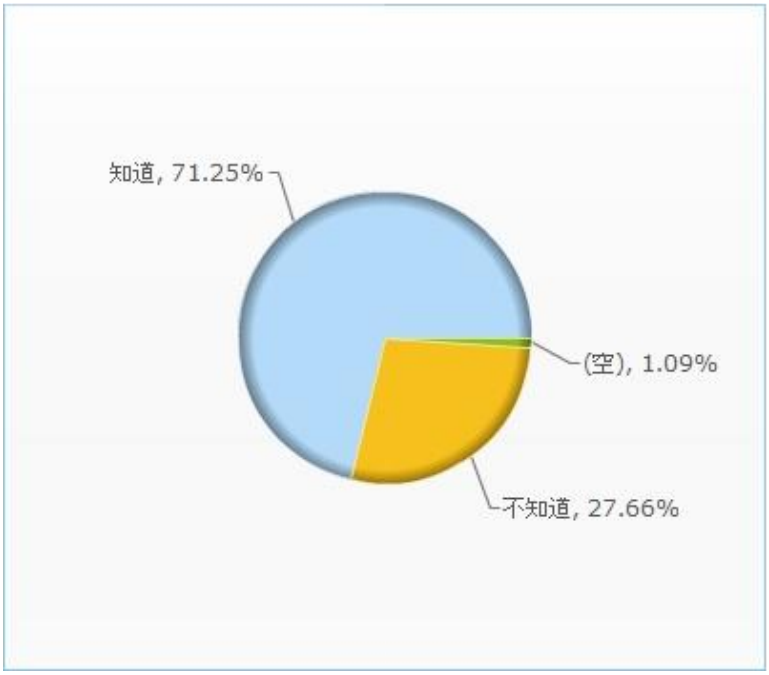
三、 问卷回收情况统计

本次调查面向中科院各研究科研人员，回收样本总数为 640 份，通过分析填写自己所在研究所的问卷，来源于 65 个研究所，各研究所的答题参与人数如下图所示（考虑篇幅，图中略去了未填写来源研究所和参与人数<5 的情况）：



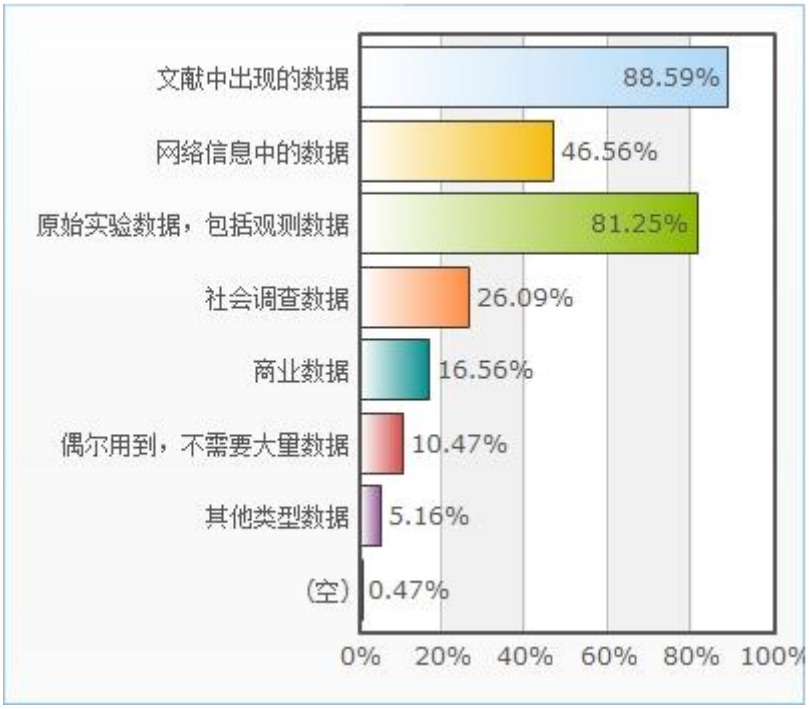
四、 调研结果分析

1、 大多数受调查者知道有科研人员利用数据挖掘或分析方法来发表论文



在参与此次问卷调查的科研人员中，知道有科研人员利用数据挖掘或分析方法来发表论文的占到 71.25%，同时也有将近 30%表示不了解，剩余 1%的人未填写。可见，知道有科研人员利用数据挖掘或分析方法来发表论文的科研人员占大多数。

2、 在科学研究中，科研人员最常用到的数据类型为文献中出现的数据、原始实验数据（包括观测数据）以及网络信息中的数据



对于被调查者而言，科研过程中最常用到的数据类型为文献中出现的数据和原始实验数据（包括观测数据），均占比 80%以上；其次是网络信息中的数据，占 46.56%；社会调查数

据和商业数据也有部分科研人员用到，分别占 26.09%和 16.56%；有 10%左右的科研人员表示不需要大量数据。

另外，5%的受调查者用到了其他类型的数据，类型大致包括以下几类：

（1） 统计年鉴、社会经济数据、公共服务行业数据、经济数据、产业数据、新闻广播中的统计数据；

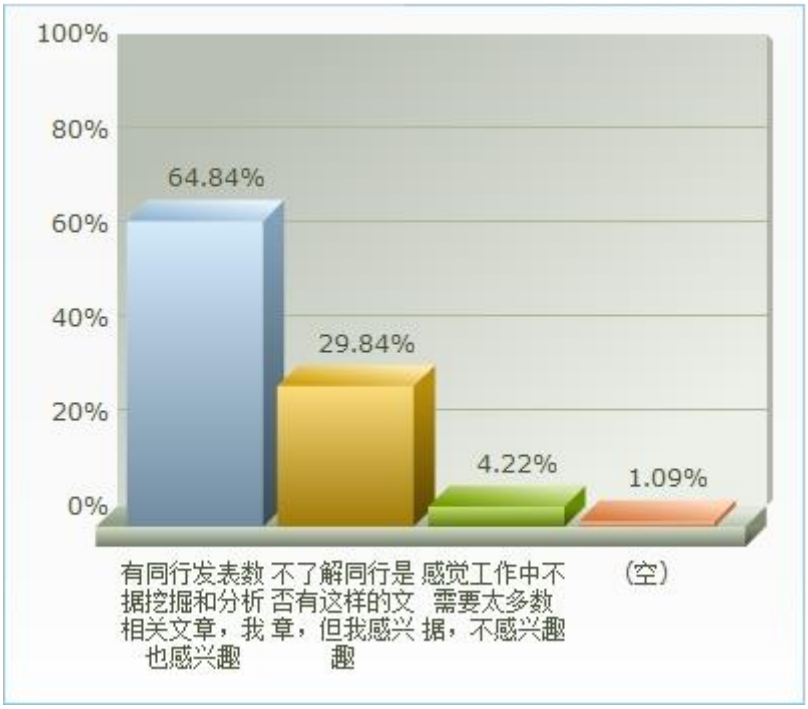
（2） 专业数据库，例如晶体结构 ICSD；

（3） 研究报告、年报；

（4） 计算模拟数据、网络开源代码数据、计算服务器计算得到的数据；

（5） 古籍文本数据。

3、多数科研人员表示，有同行发表数据挖掘和分析相关文章，并且也对这种研究方法感兴趣



近 65%的科研人员对基于大量科技数据进行科学研究的方法感兴趣；另外，近 30%的科研人员不了解同行是否有这样的文章，但是却对该研究方法感兴趣；其余人员则感觉工作中不需要太多数据，对此不感兴趣。

4、科研人员在利用大量数据时，常遇到的困难是，由于数据量庞大且缺乏规律，不知道如何整理，以及缺少高效的数据采集、处理、分析工具

在科研人员利用大量数据所面临的困难中，由于数据量庞大且缺乏规律，不知道如何整理，以及缺少高效的数据采集、处理、分析工具，这两个问题是科研人员遇到最多的困难，均占比 60%多。其次是数据来源的问题，57.66%的科研人员表示不知道从哪儿找到或获取所需要的数据，找不到合适的数据提供方；42.66%的受调查者由于不具备数据分析处理技术，因此在利用数据过程中遇到困难；36.25%的受调查者表示，虽然有利用数据的想法，但是还没想好，找不到合适的机构进行咨询。另外，还有 10.94%的受调查者反映，与 IT 公司的技术人员沟通不畅，或者他们收费太高，造成了他们在利用数据时的困难。

其余有少量科研人员还遇到了其他方面的困难，例如：

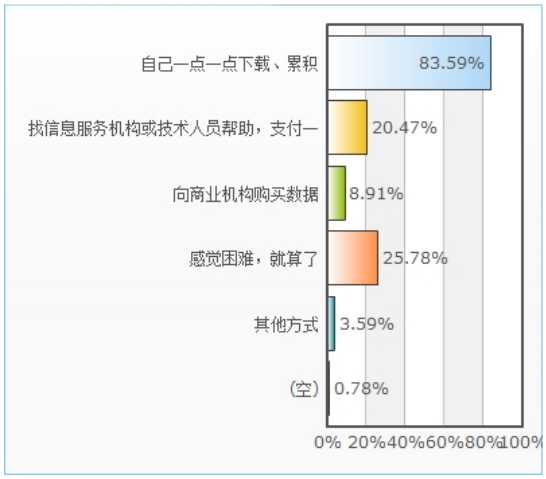
（1） 数据的准确性、可靠性、可视化等质量、功能问题；

（2） 缺乏公认的专业数据库，由于专业力量和资助力度不够，国内机构很难实现对数据的高效管理和专业数据库的构建；

（3） 科研人员的时间和精力有限，一方面不知道如何全面获取数据来源，另一方面逐一收集整理的话太耗费时间，出科研成果较慢；

（4） 目前一些数据利用被设置了人为障碍，越来越趋向一种商业行为，有些数据按照单个收费，收费特别高。

5、绝大多数科研人员通常就只是依靠人工下载、积累的方式获取数据



高达 83.59%的科研人员只能依靠人工方式，一点一点下载、累积来获取数据；有 20.47%的科研人员会通过支付一定的费用，寻求信息服务机构或技术人员的帮助获取数据；不到10%的受调查者会选择向商业机构购买数据；另外，有 26.32%的受调查者则表示感觉困难，就放弃了获取数据。

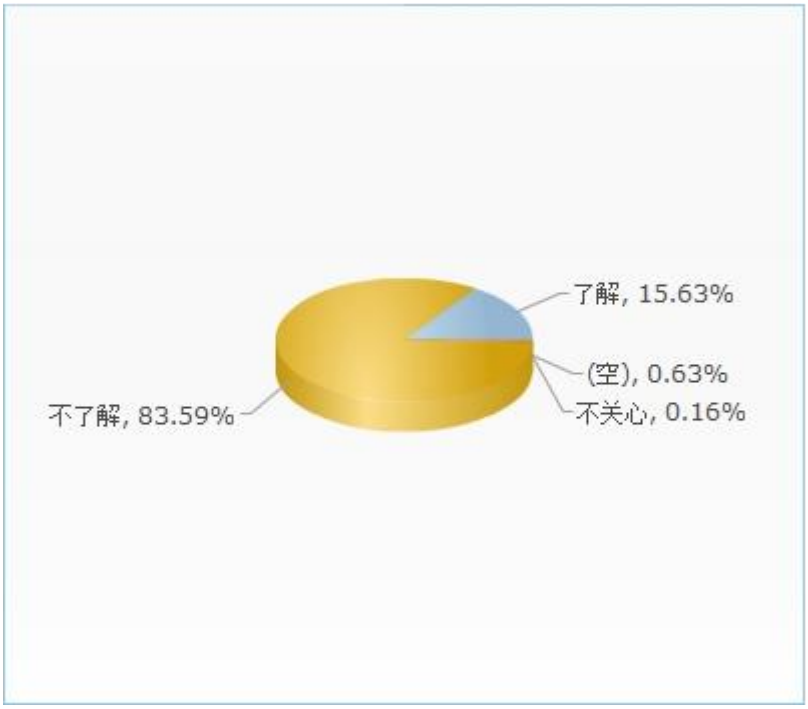
少量科研人员采用了其他方式来获取数据，包括：

（1） 通过自己所在单位或文献情报中心搜集数据；

（2） 自己通过实验、仪器、计算、编写代码获得；

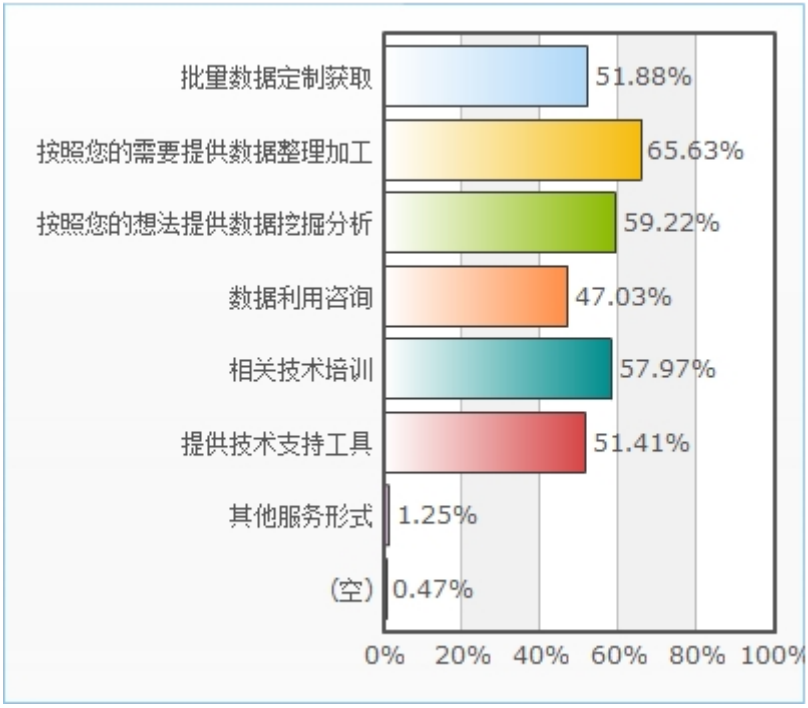
（3） 自己通过一些专业网站、国内外数据库等数据源获得。

6、大多数人不了解中科院文献情报中心可以提供数据利用的相关服务



83.59%的科研人员不了解中科院文献情报中心可以提供数据利用的相关服务，15.63%的科研人员了解这项服务，其余受调查者则不关心或未填写。因此，大多数人并不了解中科院文献情报中心正在提供的数据利用服务。

7、科研人员最希望的数据服务形式为按照其需求，提供数据整理加工和数据挖掘分析的服务



科研人员对数据服务形式的各需求所占比例较为一致，占比最多的为按照科研人员的想法和需要，提供数据整理加工和数据挖掘分析，分别为 65.63%和 59.22%；其次，对相关技术培训和技术支持工具的需求也均占到 50%多；批量数据定制获取和数据利用咨询方面的服务需求也占至 50%左右。

另外，部分科研人员还提出了其他服务方式，例如：

（1） 提供原始数据；

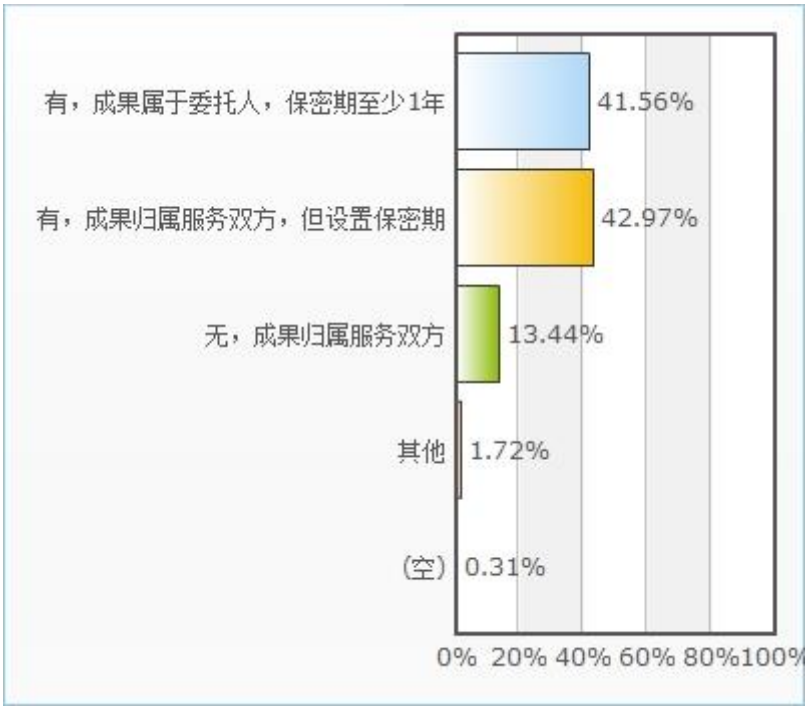
（2） 提供专业数据库等公认的数据源；

（3） 提供邮件定制等数据推送服务。

8、对于数据获取服务，受调查者最希望的合作方式为面向中科院科研人员提供非赢利性服务

科研人员期望的数据获取服务合作方式，所占比例较高的为面向中科院科研人员提供非赢利性服务（79.69%），以及先免费试点，然后逐步合理收费（63.13%）。其次是根据数据量计算服务费用，占到 32.19%，还有 5.31%的科研人员希望通过收取固定服务费用的方式来合作。另外，少量受调查者还提出了其他方式，例如，提供类似手机话费套餐的按需选择定制方式，以及在文献中提及或重点致谢的非赢利方式。

9、科研人员普遍对数据服务成果的保密及期限有要求，但对于成果的归属问题看法不同



根据数据统计结果，科研人员普遍对数据服务成果的保密及期限有要求，然而其中，42.97%认为成果应归属服务双方，但需设置保密期，另外 41.56%认为成果属于委托人，保密期至少 1 年。13.44%的受调查者对数据服务成果的保密及期限无要求，认为成果归属服务双方。另外，还有部分科研人员提出了其他要求，包括：

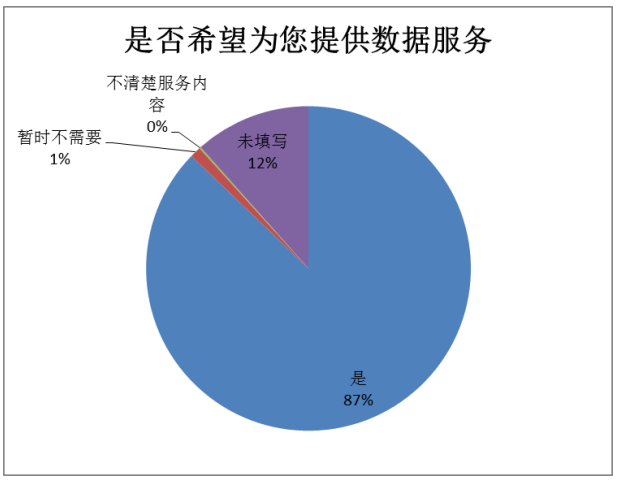
（1） 如果付费的话，成果全部属于委托人；如果是公益性质的，则归属于双方；

（2） 不设置保密；

（3） 不同的领域应设置不同的管理方法；

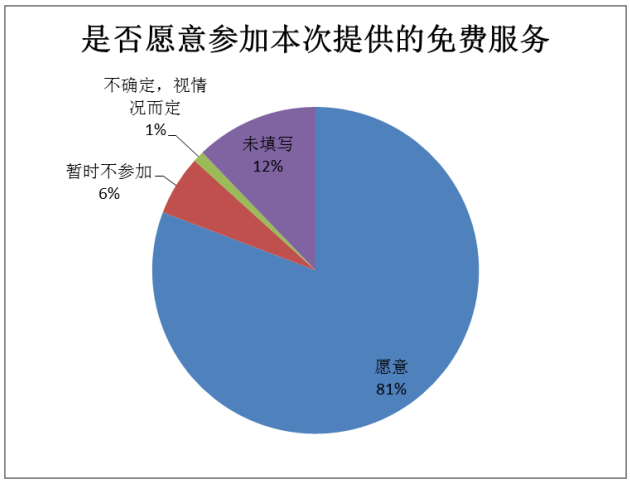
（4） 通过双方签订合同，约定相关的保密事宜。

10、中科院研究所大多数科研人员希望能够提供数据服务



在参与问卷调查的研究人员中，558 位希望能够为其提供数据服务，占 87%；只有 7 位表示暂时不需要此服务；另外，74 位未填写，1 位不清楚服务内容。可见，中科院研究所大多数科研人员对中科院文献情报中心提供的数据服务有意愿。

11、中科院研究所大多数科研人员愿意参加本次提供的免费服务



本次问卷调查将根据被调查者的领域和需求，筛选并提供两个免费服务席位。有 81%的科研人员表示愿意参加本次提供的免费服务，6%暂时不参加，1%则不确定，视情况而定，12%未填写。可见，中科院研究所大多数科研人员愿意参加本次提供的免费服务。

12、科研人员对数据服务的期待

科研人员对数据服务的期待大致可分为以下方面：

（1） 对大量数据的基础性处理，如统计、分类、筛选等数据预处理；

（2） 对特定领域文献进行综合的数据分析和文本挖掘；

（3） 抓取领域专业网站或数据库等特定数据源的数据并提供数据处理、加工、定制服务，例如将统计年鉴数据加工成可查询、方便检索的指标；

（4） 将具体较抽象的查询需求，逐步缩小范围直到检索到信息点，并书写出调研报告；

（5） 提供开放数据源、数据分析工具或软件、开放接口、开源算法；

（6） 提供大数据服务方面的介绍讲座，相关数据处理和分析软件或工具的培训、指导。