

大数据时代的新闻学

清华大学国际传播研究中心主任、教授、博士生导师 李希光

清华大学国际传播研究中心博士生 张小娅

欧洲新闻学中心（European Journalism Centre）和开放知识基金会（Open Knowledge Foundation）共同开发了一本《数据新闻学手册》（The Data Journalism Handbook），并为全球用户提供免费下载和使用，旨在推动数据新闻学的发展。本文结合中国的案例，重点介绍这本书的主要内容。

数据新闻作为在数据时代下精确新闻学的发展，其生产过程更为精细化，涉猎的知识技能除了传统的新闻音视频、图片与文字的制作，更涵盖社科研究方法、计算机数据抓取/搜集、处理、可视化，平面/交互设计、网络编程、新媒体艺术等多个领域。数据新闻是在多学科技术手段下，把庞杂的数据集中不同变量的复杂关系及其与整个社会发展的关系，用视觉语言向公众展示，以这种更客观更友好的报道方式激发公众对公共事务的探讨与参与，这种跨学科跨领域的新闻生产方式已经被如《纽约时报》、《卫报》等几大国际主流媒体所采纳，欧洲新闻中心、Google 等等机构也从 2011 年开始，举办了各种有关数据新闻的全球性学术会议与竞赛活动。

陶氏 (Tow Foundation) 基金会与约翰·奈特基金会 (John S. and James L. Knight Foundation) 于 2012 年 4 月 30 日宣布¹，他们将提供 20 亿美元来资助哥伦比亚大学的新闻学研究生院，旨在支持哥伦比亚大学数字新闻学陶氏中心的研究工作。按照赞助方的解释，这项新研究专注于数字新闻的以下三方面：

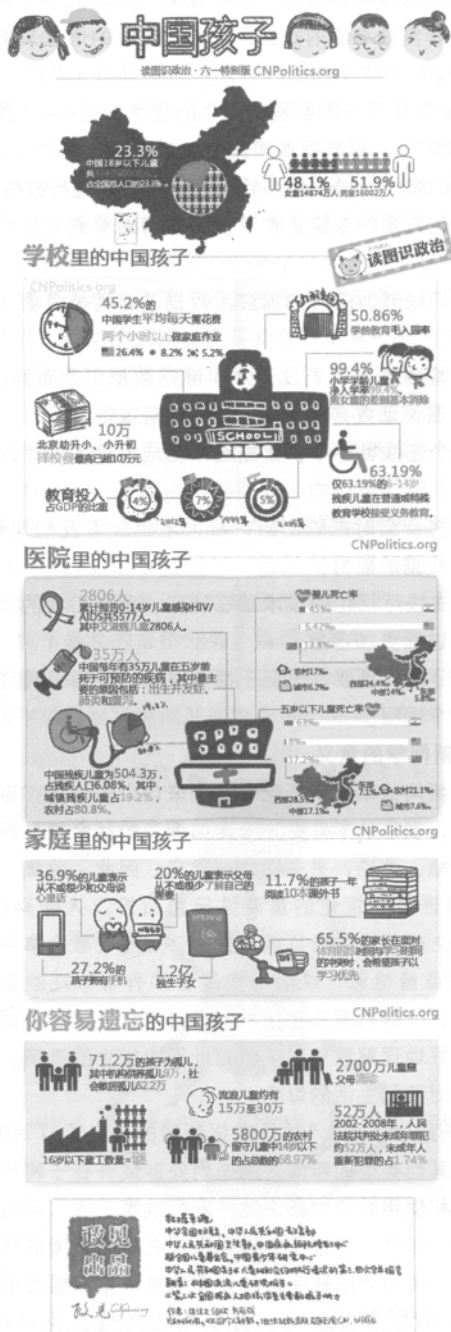
a.影响: 衡量新的实践及工具如何影响听众及新闻编辑室资源。

b.新闻的透明度：关注公共数据——哪些是可用的，哪些不是；哪些是有用并与人民的生活息息相关的；

c.数据形象化：衡量哪种形式在传达信息与吸引读者方面最有效。

哥伦比亚大学数字新闻陶氏中心主管 Emily Bell 在一篇文章中写道：“大部分新闻编辑室仍不了解数字科学的发展前沿、信息传播，及其对信息使用者的影响。我们想倡导那些具有这些领域相关技术的人才，以及对新闻学充满热情并具有相关知识的人才进行研究或提出设想。这不仅有利于解读这个世界，还能为新闻学在这个复杂多变的领域里提供指导……这项研究的目标是对新闻业以及新闻学习产生影响，广泛的、直接的影响。”²

在中国，数据新闻的发展方兴未艾，目前一些新闻传播院校已经开设相关课程，业界也有网易数读、政见CNPolitics³等，还有一些平面媒体、商业机构与个人，都在对数据新闻进行探索与尝试。

图 2-6: 政见 CNPolitics 的六一儿童节专题报告⁴

1.什么是数据新闻学

数据新闻学 (data journalism) 或称数据驱动的新闻学 (data driven journalism), 被认为是计算传播学的一个具体应用, 通过挖掘和展示数据背后的关联与模式, 和丰富的、具有互动性的可视化, 数据新闻学成为新闻学的新疆

域和应用范例，并作为一门新的新闻分支进入主流媒体。比如，在猪流感疫情爆发时，每天都有从各地传来的最新数据，《卫报》网站的数据博客（Data Blog）设计了一幅猪流感疫情的互动地图，展示世界各国的疫情进展。

任职于斯坦福大学的 Geoff McGhee 教授是一位以多媒体和信息图标为专长的记者，2009–2010 年他在约翰·奈特新闻奖学金支持下开始研究数据可视化。他认为，现在的新闻越来越多地和数据、数字有关，媒体的责任是如何向公众解释复杂难懂的数据，既给予足够的信息又不置于危言耸听。

McGhee 制作了一则数据新闻学的教学视频——“数据时代的新闻学”⁵，对数据新闻做出如下描述：

- 数据的爆炸式增长使得我们需要工具来进行分析；
- 可视化方面的专家正在开发工具帮助普通人更好地理解数据；
- 记者们则努力应对如何应用数据使新闻报道更加有说服力；
- 有经验的数据图表设计师能够把数据引入新闻学，但他们依然在论证数据对于概念诠释的有效性；
- 在一个连线世界中，数据越来越成为个人表达的载体；
- 数据将会实时推陈出新，极大地挑战着我们理解、分析和展示数据的能力；
- 创建在线可视化的技术正在转变，而新工具的出现将会使这个过程更加容易；
- 数据分析的重要性不亚于视觉展示，现有工具可以帮助实现这个过程。

2. 数据新闻学的意义

当下，新闻故事不断涌现，它们来自众多信源：目击者以及博客，发生的事件在一个浩如烟海的社会关系网中被过滤、评级、评论，更多则是被遗忘。因此，收集、过滤并将信息进行可视化的重要性日益凸显。人际网络、人—物网络之中使用的语言就是数据，那些在单个事例中无关紧要的微量信息，从全局角度看却有着非凡的重要性。现如今，一群具有开创精神的记者已经开始展示如何利用数据更好地理解我们当下所处世界中发生的事情，以及这些事件对我们生活的影响。

《数据新闻学手册》⁶ 关于数据新闻的定义这样写道：数据新闻与其他类型的新闻区别在何处？或许在于将传统的新闻敏感和使用数字信息讲述一则好故事的能力相结合而带来新的可能性，这些可能性会出现在新闻报道的任何一个阶段——使用电脑程序自动处理信息收集和组合的过程，而这些信息来自政府、公安局和其他公民机构。（Bradshaw, 2011）

数据新闻能够帮助记者使用数据图表讲述一个错综复杂的故事，比如，Hans Rosling⁷ 使用 Gapminder 软件将世界贫困进行可视化处理，吸引了来自全世界的关注；David McCandless⁸ 广受欢迎的大量数据提取——比如政府开支的背景资料，或是冰岛火山造成和阻拦的污染物——

显示出明确而清晰的设计的重要性。

数据新闻还可以帮助解释新闻事件和个人之间的关联，比如 BBC 和《金融时报》定期制作关于财政预算的互动性报道（观众可以发现预算对于自身的影响），它还可以推动新闻采访的过程，如同《卫报》成功使用数据博客分享数据、背景和议题。

数据可以成为数据新闻的信源，或是讲述新闻故事所使用的工具，也可以两者兼得。如同其他信源，数据应该接受怀疑和质疑；如同其他工具，我们应该意识到它如何塑造并限制利用数据生产出的新闻报道。

《数据新闻手册》的作者认为，通过数据的使用，记者工作的重点从第一个报道者转化成为特定事件的影响的阐释者。话题的范围宽且远：下一次正在酝酿的金融危机，我们使用产品背后的经济学，资金的滥用和政治决策的失误，显示一些抽象的社会问题，比如失业。使用数据能够将抽象概念转化为普通人可以理解并且会涉及的事物，记者们还可以分析复杂局面中的动态关系，比如骚乱和政治辩论，显示其中的谬论，帮助人们寻找复杂问题的解决方案。（Lorenz, 2011）

此外，深入的数据新闻可提供更深刻的观点。如今，编辑室数量削减，大多数记者希望改行进入公关行业，数据记者或是数据科学家已经成为相当抢手的员工——不只在媒体行业。全世界的公司和机构都在寻找“意义制造者”，即那些掌握挖掘数据、将其转换为有形信息的专家。

数据大有希望，这激发了编辑室的兴趣，促使他们寻找新型的记者。对于自由撰稿人来说，熟练的数据操作同样提供了通往新职位和稳定收入的道路。不妨这样思考：与其雇佣记者用低价值内容迅速填充版面，不如使用数据创造出对交互式内容的需求，而实现后者的唯一渠道是花费一整个星期研究一个问题，这在媒体行业中的许多领域都广受欢迎。

3. 数据新闻学的成功案例

(1) BBC

BBC 网络新闻的数据新闻工作团队包括 20 位左右记者、设计师和程序开发员。网络新闻的编辑 Bella Hurrell 和 Andrew Leimdorfer (2011) 这样介绍 BBC 对数据新闻的应用：“在新闻网站上，我们已经使用超过十年的数据，为我们的用户提供服务 and 工具。”⁹

1999 年，BBC 使用政府公布的年度数据首次发布了学校排名表，只要输入邮政编码，读者可以找到当地的学校以及一系列指标，以对其进行比较。教育记者也与该项目的开发团队合作，在发布新闻之前从这里搜索数据。当 BBC 开始这项服务时，没有官方网站可以为公众提供相关信息。而今，英国教育部已建立起自己的比较服务，BBC 也转向侧重从数据中挖掘新闻。最近的例子是，BBC 展示出一个大数据集，使用来自警方的数据记录每一条道路上的每例死亡报告。邮政编码搜索系统让用户能够找到过去十年中英国所有交通死亡事故的发生位置。为了让数据更加有时效性，BBC 与伦敦急救协会以及 BBC 伦敦电台和

电视台合作追踪在首都发生的车祸, BBC 对此进行网上直播, 并通过 Twitter 标示出事故发生位置。

BBC 不仅提供探索大型数据集的方法, 也成功创建简单的工具, 为那些惜时如金的用户提供零碎的个性化新闻产品, 这些人不会选择探索冗长的分析, 而是希望得到个人所需的信息。一个典型案例是 BBC 的预算计算器, 让用户能够发现, 当首相提出的国家预算生效时, 他们的生活会变得更好或更糟, 然后分享这一数字。BBC 与毕马威会计师事务所联手提供年度预算计算, 设计了一个有吸引力的界面, 并鼓励用户使用该项服务。

在开发新服务项目中, BBC 意识到与具有专业知识和时间进行调查的团队或是项目进行合作是最富成效的。BBC 的时事节目《全景》耗费数月时间与新闻调查中心合作收集政府部门的薪酬数据, 结果是一则电视纪录片和网络特别报告公布了所有的数据并根据各个部分的分析对数据进行了可视化处理。

除了和深度调查团队进行合作, 具备专业知识的记者也不可或缺。当团队中的商业报道记者分析了政府提出的削减开支数据, 他的结论是政府提交的数据有点夸大其词, 结果是一篇辅以清晰可视化的独家报道《理解数据》, 该报道赢得了皇家统计学会奖。

(2) 《芝加哥论坛报》

《芝加哥论坛报》的新闻应用程序团队是一群在编辑部工作的黑客, 他们与编辑和记者密切合作, 帮助他们研究并报道新闻、图解报道, 为芝加哥市民制作常用不衰的网络资源。

该报的新闻应用程序编辑 Brian Boyer (2011) 这样介绍他们的工作: “所有关于新闻应用程序的创意都来自编辑室中的记者和编辑, 这是我们和其他编辑室最大的不同。我们之间建立起了非常牢固的个人以及专业关系, 当他们(编辑和记者)获得数据, 会向我们提供想法。”程序员的工作主要是辅助记者, 帮助他们挖掘数据、将 PDF 文件转化为电子数据表、对网站进行屏幕搜刮 (screen scrape), 等等。程序员很乐意提供这些工具, 因为这能让他们实时处理编辑室中正在进行的数据工作, 将其转化为应用程序——一张地图、一个表格, 或是一个更大规模的网站。此前, 程序员将程序和新闻报道链接起来, 没有收到太大成效, 现在程序占据了网站最显著的位置, 和报道相辅相成。“我们工作的目的是产生影响——他人的生活、法律制度, 使得政治家们承担责任。”让读者们在阅读报道后有所行动, 对自己的生活提出问题, 诸如我们的家人是否安全, 我们的孩子接受的教育是否合适? 这些程序员的工作就是让读者们能在数据中找到属于自己的故事。

(3) 《卫报》的数据博客

《卫报》数据博客曾是一个小博客, 用来为新闻报道提供完整的数据包。它的组成包括: 一页头版¹⁰、世界各国政府和全球发展数据检索、来自全球和报社的绘图艺术家设计的可视化以及用来深入分析公共开支数据的工具。

每一天, 数据博客的作者都使用 Google 电子数据表来分享新闻工作背后的完整数据, 对其进行可视化和分析, 然后为报纸和网站提供故事。

《卫报》博客编辑 Simon Rogers (2011) 认为, 最近几年公共数据的开放令人难以置信, 为数据新闻提供巨大的发展机遇。在英国普选中, 每个重要政党都致力于数据的公开和透明, 向所有人开放电子数据库。随着网络生成越来越多的数据, 读者对新闻背后的原始事实的兴趣也与日俱增

《卫报》对数据新闻最早的应用发生在 2011 年英国伦敦暴乱事件中。他们成立了“解读暴乱”工作团队, 在特别项目编辑 Paul Lewis 的指挥下, 广泛使用大量数据, 帮助读者更好地理解事态进展和背后原因。与此同时, 《卫报》还与学界合作, 邀请曼彻斯特大学的 Rob Procter 带领的学术团队一起研究社交媒体在暴乱中的作用, 后者一共分析了 260 万条关于暴乱的推特, 观察谣言如何在推特上传播、不同的用户在宣传和散布信息中的功能, 以确定推特和其他组织是否煽动了暴乱。

《卫报》的“解读暴乱”数据团队使用简单的地图, 显示暴乱发生地点的贫困程度, 让“暴乱与贫困没有关系”的主流政治话语不攻自破。他们还制作了一段视频, 将暴乱发生地和参与群众的家庭住址联系起来, 显示出“暴乱通勤路线”, 建模预测暴乱者最有可能采取的路线。此外, 他们还展示出网络流言的传播途径。研究者按照话题将关于暴乱的推特分类, 编码为重复、驳斥、质疑和评论, 然后进行可视化处理。该研究发现了主流媒体在流言传播中的明显角色以及推特在矫正流言中的作用。

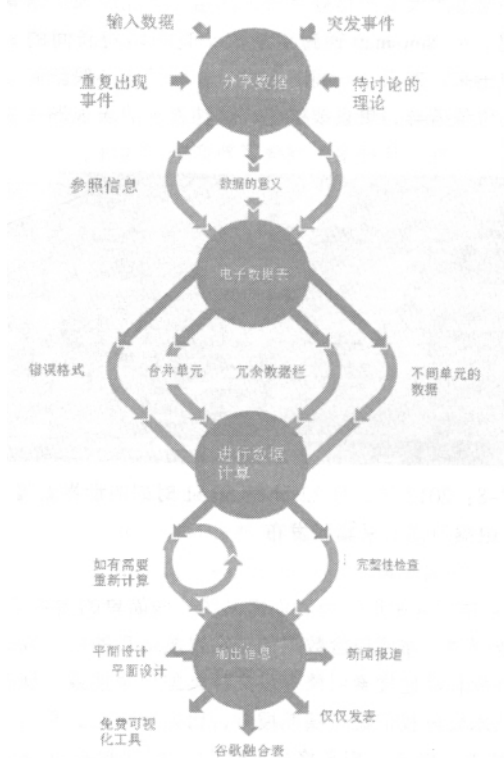


图 2-7: 《卫报》数据博客的背后¹¹

(4) 新闻地图

快节奏的数据新闻周期需要记者们在海量网络信息中寻找所需之物。新闻工具层出不穷，但是极少能像 Newsmap¹² 一样积累起众多的忠实用户。

新闻地图是 Marcos Weskamp (现任 flipboard¹³ 的总设计师) 在 2004 年设立的互动新闻门户，使用树映射的方法将谷歌新闻聚合的实时新闻数据进行可视化，页面采用 Flash 构建。与大多新闻聚合站点堆砌的列表排列形式不同，Newsmap 采用了一种信息视觉化的手法将新闻展现出来，正如其名，它像一张新闻地图，极具视觉冲击力。

新闻地图的特点：

- 高效：只在一秒钟内处理全世界主流媒体的最新发布新闻，内容还可以根据用户在搜索栏中输入的关键词而随时更改。

- 操作简单，Newsmap 采用了若干个大小不一的色块，每个色块上只显示新闻标题，当把鼠标悬停在色块上，便会显示新闻预览。色块是构成 Newsmap 最基本的元素。提高读者的视觉体验。

- 定制方便，用户可以注册一个账号，然后在系统中记录下自己的搜索偏好，即可处理根据国家、类型、发表时间显示新闻数据。

从表面上看，Newsmap 只是在表现形式上不同于一般的新闻站点，但表现形式的差异归根于对新闻重要性与时效性的认知。几乎所有的新闻站点都是依照时间顺序进行排列，即后来居上，突出的例子就是门户的滚动新闻。在速度即生命的网际传播中，这样做是为了保障新闻的时效性，但与此同时，突出新闻重要性的手段就乏善可陈，一般都是通过放大新闻标题字号，加粗、高亮新闻标题等手段实现。而 Newsmap 通过色块大小直观地将新闻的重要性反映出来，兼顾了新闻时效性，体现了与一般新闻站点核心的价值差异，更重要的是，这种直观的表现形式显得灵动活泼，在一定程度上增强了新闻的可读性。

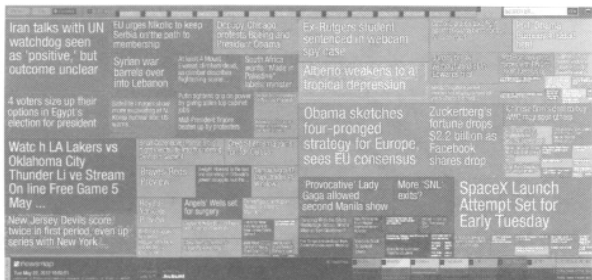


图 2-8：2012 年 5 月 22 日 16:50:51 时刻的世界新闻¹⁴

4. 数据新闻的采集和发布

(1) 数据收集

《数据新闻手册》为我们提供了一些简单的搜索建议。搜索数据时，需要包含数据的内容以及信息的格式或是来源。谷歌和其他搜索引擎提供文件类型分类搜索。现在的网络技术允许我们进行精准搜索，比如电子数据表（在搜索时附加“格式：XLS 格式：CSV”）、地理数据（“格式：SHP”）、数据库抽取（“格式：MDB，文件类型：SQL，文

件类型：DB”），或是 PDF 文件（格式：PDF）。

另一种方式是从网络的专用数据端口、数据中心以及其他数据站点获得数据：

- 官方数据门户：越来越多的国家开始建立自己的数据门户，以促进公众和商业机构对政府信息的重新利用，诸如美国政府的 data.gov 和英国政府的 data.gov.uk。datacatalogs.org 提供了此类数据的全球最新索引。英国《卫报》的全球政府数据是一个元搜索引擎，包含许多国家的政府数据分类目录。

- data.hub 是由开放知识基金会运行的数据资源站点，便于查询、分享、并重新使用那些已公布的数据。

- scraperwiki 是一个网络工具，可以更方便地提取碎片化的数据并在其他程序中重新利用，或从记者和研究人员那里检索数据。绝大多数数据是公开的，可以再次使用的。

- 世界银行和联合国的数据门户提供了所有国家的高水平指标，通常是多年的数据。

- 出现了一些致力于数据销售和再销售的创业公司，包括 Buzzdata——私人数据和公共数据包分享及合作以及数据商店比如 Inforchimps 和 DataMarket。

- Datacouch：这是一个上传、提炼、分享和可视化处理数据的地方。

- Freebase 是谷歌的一家子总司，由开放数据的爱好者们组成，提供人、地点和事物的实体图。

- 研究数据：国家和学科的数据集合，不胜枚举，比如英国数据档案。

网络论坛是搜索数据的另一去处。GetTheData 是一个问答网站，包含各种关于数据的问题，比如如何寻找关于特定题目、如何从一个指定数据来源获取信息、可视化工具、数据清理或转变为可操作的格式。数据记者还可以使用 Data Driven Journalism List 和 NICAR-L 的邮件列表，列表中包括了众多数据记者和精通计算机辅助报道 (CAR) 的极客。除此之外，程序设计员们组成了一个数量急剧增长的国际草根新闻组织，拥有几十个分部，成员数以千计，来自五湖四海，他们的共同目的是建立涵盖数据记者和技术专家的工作网络，重新思考新闻和信息的未来发展方向。

(2) 数据呈现

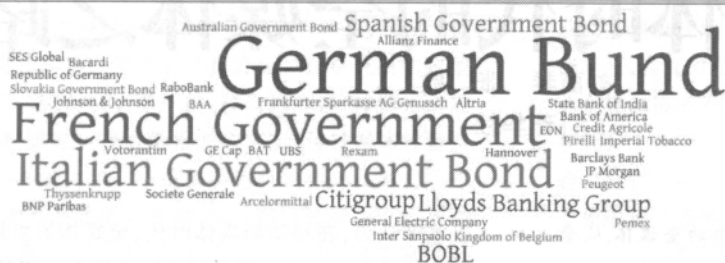
从原始数据配上新闻故事，到创造美丽的可视化和交互式网络应用程序，向公众展示数据有很多不同的方式。有的时候，数据讲述新闻的效果胜过文字或照片，这就是新闻应用和数据可视化在编辑室里为何如此受人瞩目的原因。新闻工具和技术的大丰收（通常是免费）激发了人们的兴趣，它可以让最不善于技术的记者们将数据转化为视觉故事。

像谷歌融合表，Many Eyes，Tableau，Dipity 和其他工具，让用户创建地图、图表、图形等数据图形变得易如反掌，此前只有专家才可以完成这些任务。记者所面对的问题则是是否应该将数据包可视化。

Visualizations: Blockbusters - Top holdings of the biggest European bond funds

Uploaded by: chensale
Created at: May 14 2012
Tags: economics europe investment funds

Edit Language Font Layout Color

图 2-9: 应用 Many Eyes 制作的欧洲公债基金主要持有者¹⁵

《西雅图时报》的 Cheryl Phillips (2011) 介绍说, 数据记者们经常通过可视化处理将数据嵌入新闻, 让读者可以轻松下载数据集, 在可视化程序中互动或是利用数据本身挖掘新闻背后的更多事实。这样的新闻向读者公开数据, 供批评者和更多感兴趣的人使用, 显示出数据记者和编辑工作的透明度, 也可以从批评者和读者那里获得更多建议, 这些对于提高新闻质量十分重要。

《纽约时报》研发团队的数据设计师 Jer Thorp 的观点是, 关于大数据的很多讨论遗漏了一项: 人性面。人们大多把数据视为分离的、自由流动的数字, 而忽略它们其实是对 (通常是很有人性的) 真实事物的测量。数据和真实的人、真实的生活紧密相连, 数据专家必须要思考生产数据的真实世界。到目前为止, 位置数据的使用者都是第三方——程序开发员、知名品牌和广告公司。第二方 (电信商和设备管理者) 拥有这些数据, 而第一方, 即我们每个人既无法得到数据也无法支配这些信息。《纽约时报》的研发团队推出了一个叫 Openpaths.cc 的原型设计, 允许公众探索自己的位置数据, 并亲身体验数据拥有者的概念, 毕竟人们应该对和他们自身生活和经历密切相关的数据有一定的控制权。

新闻应用程序允许透过新闻故事观察背后数据的通道, 它们既可能是可搜索的数据库, 还可以是漂亮的可视化产品。无论采用怎样的形式, 新闻应用程序的主旨是鼓励读者和数据互动、发掘数据的意义, 比如查询所处地区的犯罪趋势、社区医生的安全执业记录, 或是为他们选出的候选人的政绩。■

注释:

1 <http://www.journalism.columbia.edu/news/674>

2 <http://www.knightfoundation.org/blogs/knightblog/2012/4/30/emily-bell-how-new-research-effort-will-help-newsrooms-determine-whats-next/>

3 <http://cnpolitics.org/>

4 资料来源: <http://cnpolitics.org/2012/06/china-kids/>

5 <http://datajournalism.stanford.edu/>

6 <http://datajournalismhandbook.org/1.0/en/>

数据新闻手册是针对对数据新闻这一新兴领域的爱好

者所出版的一本免费开放的参考书。这本书最早开始于 2011 年在伦敦的 Mozilla Festival 48 小时工作坊, 尔后通过来自澳大利亚广播集团、BBC、芝加哥论坛报、德国之声、卫报、金融时报、赫尔辛基日报、阿根廷 La Nacion 报、纽约时报、美国

在线新闻 ProPublica、华盛顿邮报、芝加哥论坛报、世界之路报、威尔士在线、Zeit 在线等等的诸多数据新闻领域的领先倡导者与资深专家网络协作方式编写而成。数据新闻手册是由欧洲新闻学中心和开放知识基金会倡导, 由 O'Reilly 出版并且在 CCAS 协议下在线免费阅读。本书目前仅有电子版, 各个章节由不同作者完成, 参考书目中引用的格式为: 本章节作者姓名, 章节题目, 手册名称, 本章节网页地址。

7 汉斯·罗斯林 (Hans Rosling 1948 年 7 月 27 日) 是卡罗琳学院的国际卫生学教授, 并担任 Gapminder 基金会董事长, 该基金会开发了 Trendalyzer 软件, 具有把统计数据图形化的特点, 方便人们理解数据资料。

8 大卫·麦克坎德雷斯 (David McCandless) 曾为《英国卫报》、《连线》、《独立报》等刊物撰稿, 擅长以简洁精美的图像展现复杂、抽象或分散的资讯, 并将不同的数据组合, 展现其中的联系和模式。大卫认为, 数据可视化不仅是在信息丛林中找到方向的最好方法, 还能帮助人们并发现全新的视角。他的新作《信息是美丽的》(Information is beautiful) 以其擅长的可视化数据描绘了当今世界的各个方面。

9 大卫·麦克坎德雷斯 (David McCandless) 曾为《英国卫报》、《连线》、《独立报》等刊物撰稿, 擅长以简洁精美的图像展现复杂、抽象或分散的资讯, 并将不同的数据组合, 展现其中的联系和模式。大卫认为, 数据可视化不仅是在信息丛林中找到方向的最好方法, 还能帮助人们并发现全新的视角。他的新作《信息是美丽的》(Information is beautiful) 以其擅长的可视化数据描绘了当今世界的各个方面

10 www.guardian.co.uk/data

11 资料翻译自《数据新闻学手册》。

12 网址为: www.newsmap.jp

13 网址为: www.newsmap.jp

14 注: 图片来自 Newsmap 网站截图, 图中文字为新闻标题, 未翻译。

15 资料来源: <http://www-958.ibm.com/software/data/cognos/manyeyes/>。图中主体部分文字为欧洲公债基金持有者机构名称, 未翻译。