“爬虫技术”获取第三方商业数据的合规与实践

作者：杨迅 | 象玉婷 | 夏雨薇

数据是数字经济的血液，正如石油是现代工业的血液。面对互联网上广泛分布的大量数据，是否可以通过“爬虫”或其他技术手段获取第三方公开的数据，开展大数据分析或其他应用?这既是一个合规问题，也是大数据公司面临必须解决的生存问题。

近几年来，平台企业和数据分析公司之间关于获取数据的合法性的案件不断涌现，虽然，平台企业维持较高的胜诉率，但是也必须面对这样的问题：他们对数据拥有的权益的边界到底在哪里?屡败屡战、百折不挠的数据分析公司也在不断改进业务模式，探索其利用第三方数据开展大数据分析的可行性方案。

我国 11 月颁布的《反不正当竞争法修订草案征求意见稿》(“**反法草案**”)也新增第十八条，专门罗列“不正当获取或使用其他经营者的商业数据”的不正当竞争行为，似乎为规范数据的使用行为打开了大门。

本文以目前大数据公司较为普遍使用的“爬虫技术”获取平台数据为例，探讨合规的数据的获取规则。

上篇：他山之石

美国互联网产业发展的较早，数据产业也较为繁荣。相应地，与数据有关的法律规则比较宽松。

对于“爬取”数据的行为，在美国法律下，涉及一项古老的普通法概念“Trespass to Chattels”。其本意是指：侵权方故意(或，在某些法域，疏忽)干涉他人对动产的合法占有的侵权行为。而网络时代，一些法院接将计算机系统认定为“chattels”,认定未经许可的在他人计算机系统上的操作行为，形成“trespass”。根据网络财产理论，计算机动产所有人有权禁止他人与其系统进行非破坏性接触。从本质上讲，网络财产相当于将他人排除在网络相关资源之外的

如您需要了解我们的出版物，
请联系:

Publication@llinkslaw.com

权利。该权利类似于将他人排除在不动产之外的权利。比如在 eBay Inc. vs. Bidder Edge Inc.1 案件中，法院授予了 eBay Inc.禁止 Bidder Edge Inc.通过技术手段获取其数据的禁令，法院认为 Bidder Edge Inc.获取 eBay Inc.的数据虽然没有在现实中损害 eBay Inc.的权利，但是如果允许它未经 eBay Inc.许可获取 eBay Inc.数据，必将引起其他公司效仿，因此这样的行为可能给 eBay Inc.带来损害。在 Oyster Software Inc. vs. Forms Processing Inc.2案件中，法院甚至更进一步指出：无需证明损害，一旦第三方未经许可进入他人计算机空间，就属于侵权。但是，随后的 TicketMaster vs. Ticket.com3一案，法院有了不同的看法，在该案中，纵然 TicketMaster 表明其收集、整理购票信息花费了大量的精力，因而该些信息是有价值的，法院仍然认为 Ticket.com 通过技术手段获取该些信息，没有损害 TicketMaster 的利益，不构成侵权。

近期，美国法院对“爬取”数据采取更加开放的态度。2022 年 4 月 18 日宣布的 hiQ Labs[4], Inc. v. LinkedIn Corp.判决，重点考察信息的公开性，强调对于公开信息的可触达。美国联邦第九上诉法院认为，访问一个对外公开，皆可访问的计算机时，并不存在“未经授权或超出授权”的概念，此时访问该计算机并获取数据的行为也不构成《计算机欺诈和滥用法》所规制的违法行为。类似地，2021 年 6 月 3 日终审判决的 Van Buren v. United States[5]，美国联邦最高法院也确认，《计算机欺诈和滥用法》约束“超出授权”获取计算机数据的行为，本身有权获取数据的人即使对数据进行不当使用，也不能解释为“超出授权”。

相对而言，英国和欧洲对“爬取”数据的态度比较保守。由于欧洲有数据库保护指令(EU Database Directive)，因此关于爬取网络数据的讨论，往往和数据库权利结合在一起。一般而言，对于受保护的数据库，如未经权利人的同意下，抓取或重新利用所有的、或实质性部分的数据库数据的行为，则可能构成侵权。

作者受到数据库权利保护的前提为，数据库是独立数据、作品、材料的集合。比如，欧盟法院 2004 年 Fixtures Marketing Ltd v. Svenska Spel AB (“Fixtures 案”)[6]，与 2012 年 Freistaat Bayern v. Verlag Esterbauer GmbH[7]案(“Bayern 州案”)中认定，数据库是独立数据的集合，独立意味着数据自数据库中提取后应具备“自动的信息价值”(autonomous informative value)。Fixtures 案中足球赛程表中的日期、球队和比赛时间从赛程表中剥离出来后，对博彩客户而言有信息价值。而 Bayern 州案中原告地形图中剥离出的坐标点、特定图标对于想要制作新地图的第三方而言也是有价值的。

进一步地，数据库权利要求数据库作者对数据库的获取、验证和呈现进行了“投资贡献”。“投资贡献”必须是对数据库(例如其验证体系、搜索编排体系)而非对数据本身的贡献。上述的 Fixtures 案中主办方安排赛程时，其贡献在于生成赛程数据，而赛程数据生成后赛程表也已经自动生成，因为从生成数据到集合数据之间没有独立于数据生成的额外努力，故据此产生的数据集合——赛程表不受到数据库权利的保护[8]。

对于不受到数据库权利保护的数据，其仍可以通过合同限制用户对数据的使用(虽然该等约束并不是必然有效)。欧盟法院 2015 年在 Ryanair v PR Aviation[9](“Ryanair 案”)中确定，数据库所有者无法获得知识产权(包

[1] eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000)
[2] Oyster Software Inc. v. Forms Processing Inc., et al. N.D.Cal. Dec. 6, 2001, No. C-00-0724 JCS) 2001 WL 1736382
[3] Ticketmaster Corp. v. Tickets.com, Inc., 2003 WL 21406289, (C.D. Cal. March 7, 2003)
[4] hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180
[5] Van Buren v. United States, 141 S. Ct. 1648
[6] Fixtures Marketing, C-444/02, EU:C:2004:697
[7] Verlag Esterbauer, C-490/14, EU:C:2015:735
[8] Fixtures Marketing C-338/02 , EU:C:2004:696
[9] Ryanair, C-30/14, EU:C:2015:10

括数据库权利)保护时，其可以自行通过限制性协议约束访问者的使用。在该案中被告运营一个荷兰的价格对比网站“Wegolo”，允许登录该网站的用户对廉价航空的机票价格进行对比，并直接通过 Wegolo 下单机票。其通过自动化方式从线上获取航空公司的票价，其中也包括从原告网页所链接的数据集。而一般用户如需访问原告网页的，需要勾选原告网页的使用协议，其中明确禁止了数据爬取行为。欧盟法院认为，在与国内法律不冲突的情况下，不受 Directive 96/9/EC 保护的数据库的作者，可以对第三方使用数据库的行为设定合同限制。不过，该案件在回到荷兰的国内程序后，海牙上诉法院认为合同订立需要具备“要约”与“承诺”，而在本案中被告并没有做出愿意遵守原告网站服务条款的“承诺”。因为被告在访问一个数据完全公开、完全免费、对所有公众可及的网站，在一个客观理性的人看来，访问该等性质的网站并收集数据的行为随机且常见，不能被认为是被告自愿受网站服务条款的约束的特定承诺。该决定也被荷兰最高法院认可[10]。

下篇：我国的合规实践

目前，我国法律并没有对使用爬虫技术获取第三方数据做出直接规定。政策上，一方面，出于鼓励大数据产业发展和打破数据孤岛考虑，我国鼓励数据的流通，因此爬虫技术的使用存在积极价值；另一方面，纵容爬虫技术的滥用也确实更容易滋生“搭便车”现象。

反法草案指出了四类不正当获取数据的行为:(一)以盗窃、胁迫、欺诈、电子侵入等方式，破坏技术管理措施，不正当获取其他经营者的商业数据，不合理地增加其他经营者的运营成本、影响其他经营者的正常经营;(二)违反约定或者合理、正当的数据抓取协议，获取和使用他人商业数据，并足以实质性替代其他经营者提供的相关产品或者服务; (三)披露、转让或者使用以不正当手段获取的其他经营者的商业数据，并足以实质性替代其他经营者提供的相关产品或者服务; (四)以违反诚实信用和商业道德的其他方式不正当获取和使用他人商业数据，严重损害其他经营者和消费者的合法权益，扰乱市场公平竞争秩序。

我国法院在司法判决中往往以技术中立为原则，不否定爬虫技术的合法性，但就其具体使用，则严格限定边界和条件。从司法实践的角度来看，爬虫技术的合法性，需要从数据的获取方式，获取的数据内容，是否侵犯了被爬取平台及第三方合法权益，以及是否对被爬取平台的业务构成实质性替代等方面来综合分析。

(一) 数据获取的方式

数据的获取方式主要考察“爬虫”是否入侵了被爬网站的系统，还是以模拟用户方式访问。具体而言，考察以下两个因素:

(1) “爬虫行为”访问模式是否绕过了系统的安全设置

根据北京海淀法院 *(2017)京0108 民初24512 号*案件的判决:“网络爬虫等技术手段虽系自动抓取网络数据的程序或脚本，但如其遵守通用的技术规则，亦无需访问权限即可访问上述微博平

[10] Gerechtshof Den Haag 23 januari 2018, ECLI:NL:GHDHA:2018:61, Hoge Raad 27 september 2019, ECLI:NL:HR:2019:1445

台公开数据。因此，无论是通过用户浏览或网络爬虫获取该部分数据，其行为本质均相同，微梦公司在无合理理由的情形下，不应对通过用户浏览和网络爬虫等自动化程序获取数据的行为进行区别性对待”。即该法院倾向认为网络平台在无合理理由的情形下，不应对通过网络爬虫获取和用户浏览公开数据区别对待，网络平台对他人在满足合法、正当、必要的原则下抓取公开数据的行为应负有容忍义务。

相反，如果使用爬虫技术绕过安全设置的行为，窃取被访问的网站的非公开信息的，可能构成侵犯商业秘密。我国《反不正当竞争法》(“**反法**”)第九条第一项新增了以“电子侵入”方式窃取商业秘密的行为，而“电子侵入”则包括了以爬虫技术绕过被访系统的安全设置，获取非公开信息的情形。

情节严重的绕过安全设置的行为甚至可能构成刑事犯罪。根据《刑法》第二百八十五条第 2 款的规定，违反国家规定侵入计算机信息系统或者采用其他技术手段，获取该计算机信息系统中存储、处理或者传输的数据，或者对该计算机信息系统实施非法控制，构成非法获取计算机信息系统数据、非法控制计算机信息系统罪。*(2019)鲁 0213 刑初 144 号*案件中，被告人首先利用“SQL 注入漏洞”获取网站的后台管理权限，进而利用其编写的爬虫脚本程序侵入计算机信息系统，获取计算机系统内存储的大量数据，且该等数据并非在公开页面显示的数据，因此被告人的行为被认定违反了《刑法》第二百八十五条的规定，构成“非法获取计算机信息系统数据罪”。

(2) **是否会对平台服务器运行产生额外的负担**

此外，一些法院会将爬虫技术是否会给被抓取的平台服务器的正常运行产生额外负担，导致系统负载过高，进而导致平台运营的成本的增加作为认定爬取行为是否正当的因素之一(例如: *杭州互联网法院-(2021)浙 8601 民初 309 号*、*广州天河区人民法院-(2019)粤 0106 民初 38290 号*、*北京市海淀区法院-(2018)京 0108 民初 28643 号*)。最近，(2121)浙 8601 民初 309 号判决，针对使用某些爬虫技术获取数据，详细指出“绕开微信客户端，从而获得了等同于‘登录用户’的权限，同时使用自动化脚本不间断爬虫‘爬取’，异化了微信公众号去中心化的产品展示规则，会对微信公众号平台服务器造成远超正常用户访问的负担，已经妨碍、 破坏了两原告合法提供的网络产品与服务的正常运行”。

在情节严重的情况下，给被访计算机系统造成安全威胁的，甚至可能构成犯罪。根据《刑法》第二百八十六的规定，违反国家规定对计算机信息系统功能进行删除、修改、增加、干扰，造成计算机信息系统不能正常运行，后果严重的，构成破坏计算机信息系统罪。例如在*(2019)粤 0305 刑初 193 号*案件中，被告人开发的爬虫软件在 2018 年 5 月 2 日 10 时至 5 月 2 日 12 时的两小时内，以每秒 183 次的频率访问“深圳市居住证系统”，导致“深圳市居住证系统”停止运行超过 2 小时，该等爬虫使用行为被认定违反了《刑法》第二百八十六条的规定，构成“破坏计算机信息系统罪”。

(二) 获取的数据内容

从获取数据的性质看，爬取非公开信息的，可能构成违法违规。在上海市普陀区公布的数据合规第一案中，Z 公司构成违规的一项重要因素就在于其爬取了外卖平台的非公开数据。

如前述第一(1)点所述，以“电子侵入”方获取他人商业秘密的，可能构成侵犯商业秘密。反法亦设定“自愿、平等、公平、诚信的原则”，制止“扰乱市场竞争秩序，损害其他经营者或者消费者的合法权益的行为”。由此，所获取的数据是否具有独立价值是考查的重要因素。例如淘宝诉美景案(*杭州中院-(2018)浙 01 民终 7312 号*)中，法院认为：“应当区分平台对数据的投入度因素，即简单的对用户信息的转换、记录不足以使平台获得独立的权益。”；再如，2020 年 8 月的微信诉群控软件案(*杭州中院-(2020)浙 01 民终 5889 号*)中，法院认为：“应当区分数据资源整体和单一用户数据，网络平台方对于数据资源整体与单一数据个体所享有的是不同的数据权益。就平台数据资源整体而言，系平台投入了大量人力、物力，经过长期经营积累聚集而成的，该数据资源能够给平台带来商业利益与竞争优势，平台对于整体数据资源应当享有竞争权益。”

虽然目前在大量案件中，并非所有法院会就第三方获取的平台数据类型进行细化区分，但我们发现不少法院对数据流通、使用的利益平衡问题进行了特别关注，如上海知识产权法院曾在**大众点评诉百度地图案**(*上海知识产权法院-(2016)沪 73 民终 242 号*)中指出：“即使平台可以对用户发布的公开内容数据进行权益主张，但是考虑产业发展和互联网环境所具有信息共享、互联互通的特点，需要兼顾信息获取者、信息使用者和社会公众三方的利益。”

对于数据是否公开的判断，法院在很大程度上考虑公众接触数据可能受到的限制。*(2022)京 73 民终 1154 号判决书*将数据分为三类：“互联网环境下数据获取(处理)方式有三：一是对公众开放且不需要授权的数据处理；二是需要授权但已获得授权的数据处理；三是需要授权但未获得授权的数据处理(或者超出授权处理权限的处理)……数据控制者通过代码限制界定其数据可处理区域，设置用户行为规则，他人破坏或者违反代码限制而处理该数据,即构成‘未经授权’。”可见，衡量商业道德时，本身也考虑数据的公开性，至少部分取决于数据控制者以代码形式设定的边界，在多大程度上允许公众接触和使用。

(三) 是否侵犯知识产权或个人信息相关的权益

如果所爬取信息的行为构成著作权、商标权等知识产权的侵权，或者(对爬取个人信息而言)构成对个人信息的侵犯，则该种信息爬取行为则是违规的。

(四) 是否构成实质性替代

爬取数据之后的使用行为也是决定“爬虫”行为合法性的重要考察因素。一般而言，爬取的数据不得用于被爬者相竞争的商业活动。

根据《反不正当竞争法》第十二条的规定，经营者利用网络从事生产经营活动不得利用技术手段，通过影响用户选择或者其他方式，实施妨碍、破坏其他经营者合法提供的网络产品或者服务正常运行的行为。根据《关于适用中华人民共和国反不正当竞争法若干问题的解释(征求意见稿)》第二十六条规定，经营者违背诚实信用原则和商业道德，擅自使用其他经营者征得用户同意、依法收集且具有商业价值的数据，并足以**实质性替代**其他经营者提供的相关产品或服务，损害公平竞争的市场秩序的，人民法院可以依照反不正当竞争法第十二条第二款第四项予以认定。比如在前述淘宝诉美景案中，美景爬取数据用于开展与淘宝生意经类似的业务，则构成不正当竞争。

相反，如果爬取数据不用于类似业务，而是丰富了相关数据领域的生态，则不太可能构成不正当竞争。比如，在前程无忧诉上海逸橙案(*上海知识产权法院-(2019)沪 73 民终 263 号*)中，法院认为：“随着互联网市场竞争的日趋激烈，互联网市场领域的各种产品或者服务关联性和依附性不断加深，依赖甚至介入于其他经营者的产品或服务而开展经营活动本身并不会损害正常的市场秩序，相反以此而否定该行为的正当性，无疑将会挫伤创新动力。同时，被告提供的产品功能，并不违背行业惯例，可以提高工作效率，给市场主体带来便利，被告并未强制、欺骗用户使用产品功能等等，从而综合认定该行为并不具有不正当性。”

如您希望就相关问题进一步交流，请联系:

杨 迅
+86 21 3135 8799
xun.yang@llinkslaw.com

如您希望就其他问题进一步交流或有其他业务咨询需求，请随时与我们联系: **master@llinkslaw.com**

上海	北京	深圳
上海市银城中路 68 号 时代金融中心 19 楼 T: +86 21 3135 8666 F: +86 21 3135 8600	北京市朝阳区光华东里 8 号 中海广场中楼 30 层 T: +86 10 5081 3888 F: +86 10 5081 3866	深圳市南山区科苑南路 2666 号 中国华润大厦 18 楼 T: +86 755 3391 7666 F: +86 755 3391 7668
香港	伦敦	
香港中环遮打道 18 号 历山大厦 32 楼 3201 室 T: +852 2592 1978 F: +852 2868 0883	1/F, 3 More London Riverside London SE1 2RE T: +44 (0)20 3283 4337 D: +44 (0)20 3283 4323	

www.llinkslaw.com

Wechat: LlinksLaw

本土化资源 国际化视野