Web Crawling as alternative data, a regulatory perspective.

Web crawling is becoming broadly used as a source of alternative data and an investment research tool by both the sell side and by asset managers including hedge funds.

However, firms utilizing web crawling or harvesting, need to manage the associated compliance risks due to an ever growing body of regulatory deliberation associated with the practice. The courts find that operating a web-crawler is a legal and expected activity on the internet (Field v. Google), however, financial firms need to minimize the potential headline risk, as well as the potential legal costs that are associated with web crawling.

From my experience in managing alternative data units on both the buy side and sell side, there are five key dimensions to web harvesting compliance:

1. An understanding of the evolving law surrounding web harvesting;
2. Review of the terms and conditions associated with the websites crawled;
3. Control over the potential interference with harvested web sites;
4. Review of web harvesting projects;
5. Review of vendors utilized for web harvesting.

Sparsity of case precedent

There is little doubt that prominent sites are paying close attention to usage patterns and some are notorious for using legal means to stamp out non-human access to their site. Craigslist.com, for instance, has successfully sued several organizations over harvesting data from their site. Yet, research shows that fewer than 50 known web crawling cases, in total, have ever gone to court; a minuscule amount if we consider how enormously widespread the practice has become. Anecdotally, over 70% of the members of New York's statistical society meetings have engaged in web harvesting at one time or another according to an informal poll. Yet if we filter for only those cases that have a relevance to asset managers, the picture looks different still.

Relevance to the investment community

Investors seek to collect online information in order to better understand the wider trends impacting the markets. They are seldom, if ever, seeking to re-distribute the data, compete with the host, or put a significant burden on web resources. Yet, most of the current legal precedent rests on the issues which do not have a direct corollary to the investment use case. In the context of the investment community, case precedent is even thinner than in the general domain, thus increased diligence is all the more pertinent.

There are no known web crawling cases related to any category of asset management, but one known case involving financial institutions is Barclays Capital Inc. v. Theflyonthewall.com, Inc. where Barclays took issue with Theflyonthewall.com harvesting and redistributing Barclay's research reports. The case was ruled in favor of the defendant on appeal. Although it involves a financial institution, the case doesn't have many similarities to funds' usage as it involved a copyright issue.

Within the alternative data context, a somewhat relevant example case is the Internet Archive v. Suzanne Shell: Shell operated a website giving advice to alleged criminals. The Internet Archive crawled Shell's site with no commercial or competitive intent, its sole goal was to archive. Shell sued the Archive using a breach of contract (TOU), civil theft and RICO as legal ammunition. The issue was ultimately settled out of court, all the charges were dismissed except for breach of contract. It is unknown how the court would have ruled if it had been allowed to reach a verdict. This case is relevant because akin to many funds, the archive did not seek to resell or compete using the data collected. A key difference from funds' operations is that the Archive intended to make the data available to the general public. Another relevant case is Fidlar v. LPS case, where LPS, a real estate analytics company crawled Fidlar's lands records for analysis purposes. Ultimately, the case was judged in favor of LPS.

Understanding incentives

The sparsity of web crawling cases demonstrates an overall lack of motivation to take legal action by the web site operators. In part, this reluctance is due to a small likelihood of a win in court and an absence of a clear reward to the plaintiff even given a win. Thus, few companies are willing to deploy their legal resources to pursue a web crawler even if the crawler is in violation.

Tonia Klausner, an attorney partner at Wilson Sonsini Goodrich & Rosati specializing in the areas of internet data, privacy and mobile, affirms that incentives play a key role in the web data legal landscape. "The value of legal claims against web crawlers is low where the crawler does not crash or otherwise harm the website, and the crawled data is not used in competition with the website operator." Ms. Klausner says. "This is one reason why we don't see many claims being filed in court against web-crawlers, and why the claims that are filed tend to be driven by the crawlers somehow damaging the business of the data owners, whether directly or due to opportunity cost."

Many web hosts offer the general public an option to download the data via an API. While this might be a paid and restrictive option, it's recommended when available. Circumventing the API may be viewed by some hosts as lost revenue, money that they might fight for in courts.

Terms of use

Does a crawler need to comply with a Terms of Use (TOU)? In many cases, the courts find that a crawler is not bound by the TOU, especially in case of a browser wrap (see Craigslist v Mesiab, Ticketmaster v. Tickets.com, Cvent v. Eventbrite, EF Cultural Travel BV v. Zefer and Explorica and Hines v. Overstock.com). A 2009 post by D. C. Toedt III highlights more TOU related cases.

There aren't many counterexamples where the courts favored a browser wrap TOU, all with some kind of a twist. For instance, in Southwest Airlines Co. v. Boardfirst, the court ruling included a remark that Boardfirst should have abided by the TOU, but also Boardfist received and ignored two separate Cease and Desist requests before Southwest took them to court. Ignoring the Cease and Desists was "the straw" in this case, without the C&D ignore, the TOU violation by itself would likely not be enough to elicit court action.

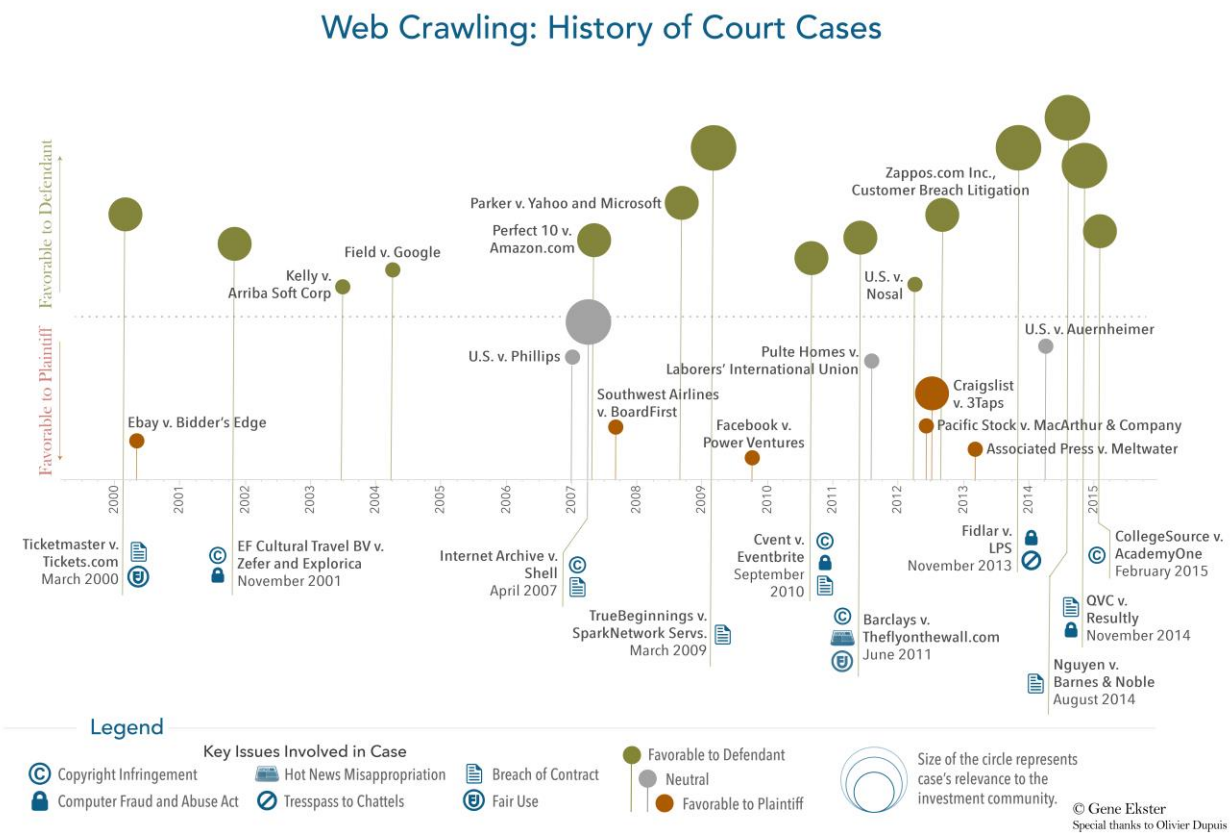
Vendor Management

If a fund is considering using a third party for web crawling, a familiar dilemma is a tradeoff between a loss of control vs. a potential legal risk offset. From a risk perspective, is it worth letting an external party build and maintain the crawlers? The answer looks to be a "yes" with caveats according to Ms. Klausner, "While the level of risk depends upon the specific circumstances, in general, purchasing crawled web data rather than obtaining it directly provides additional protection from legal claims, because claims based on web-crawling typically are asserted against the companies operating the crawler, and not their customers." Other factors to consider, is that for some funds, any downstream communication of IP (including even the identity of web resources being collected) and the loss of control over compliance practices are, understandably, unacceptable trade-offs. In addition, some may see parallels between the recent insider trading investigations into expert networks and third party web crawling solutions. However, crawling service providers are more of arm's length data vendors rather than agency-like direct connections that expert networks deliver; in the crawling case, any violations would be more akin to vendors of traditional survey administration being responsible for the interaction to their respondents. The upside of a liability offset and the fact that third parties will often have back data for popular sites are compelling reasons to outsource a web crawling operation.

Developments and ongoing diligence

While the landscape of web crawling compliance is still taking shape, recently there have been strides taken in helping to define this contentious topic. In particular, Eagle Alpha, an alternative data, research and analytics provider, has published a set of best practices related to web crawling specifically with the investment industry in mind. Further, they have held a seminar in partnership with the law firm Wilson Sonsini Goodrich & Rosati; the seminar dug into the topic covering case precedent and best practices. In addition, Integrity Research recently noted that UBS is recruiting a 'Primary Investment Research Analyst' position responsible for 'designing solutions to solve investment debates using web-harvested datasets'.

The below graphic shows the majority of the known web crawling legal cases to date, key issues involved, their outcome and their relevance to the investment industry:



Click here for underlying data

Project Review and Approval

Once guidelines are established, each web harvesting project should be evaluated prior to implementation, and reviewed on an ongoing basis. The firm should have an explicit risk assessment template to determine the project total risk based on headline/PR risk, business risk, and regulatory risk decomposition. Further, each project should receive approval from a sponsoring stakeholder and the COO or CCO.

Organizations must ensure that they stay abreast of news, cases and regulations which are added or modified because the biggest changes to the regulatory landscape are surely still to come.

Legal and compliance professionals can enhance their knowledge of data collection by reviewing the legal topics which underpin most cases thus far brought to court:

- Digital Millennium Copyright Act (DMCA)
- Computer Fraud and Abuse act (CFAA)
- Trespass to Chattels

Conclusion

Minimizing risk when web crawling can be accomplished through a combination of planned operating procedures and ongoing diligence:

- Collecting information from the web should not incentivize another organization to pursue legal action. This can be accomplished by not harming or creating an opportunity cost for that organization. For example, by ensuring that crawlers do not put excessive load on the host's web infrastructure and by avoiding directly competing with the host or reselling the data gathered. If the data host offers an API through which the data is available for purchase, even if limits are imposed, this is preferable to screen scraping.
- If the tradeoff between some loss of control is acceptable, using a third party vendor to collect web data can offset some of the liability.
- Efforts by the host to deny data collection should be respected, the few successful cases against web collection center on a cease-and-desist being ignored or an explicit circumvention of the host's technical prevention measures.

If a fund can become comfortable with the risks involved, web crawling data can be a key asset in the alternative data research arsenal.

About Gene Ekster, CFA

Gene Ekster was previously head of R&D at Point72 Asset Management (formerly SAC Capital), a Director of Data Product at 1010Data and a Senior Analyst at Majestic Research (now ITG Investment Research). Currently, Gene works with asset management firms and data providers in a consulting capacity to help integrate alternative data into the investment process. He can be reached via LinkedIn.