



The Truth and Fiction of Data Harvesting

Database publishers have a love-hate relationship with data harvesting (also known as “scraping” or “mining”). On the one hand, harvesting seems to hold the promise of drastically reduced data acquisition and updating costs. On the other hand, successful online publishers spend a lot of time and effort protecting their data from appropriation and unauthorized use from various types of harvesters, both the benign, the malicious, and the competitive.

In this white paper I am going to look at the issues I have encountered in conducting numerous harvesting efforts on behalf of a variety of publishers so that those considering harvesting or trying to protect their data from being harvested can get a sense of the current state of affairs and the options available to them. First I will take a hard look at the assumptions behind the calculations of the benefits of harvesting and then I will look at the risks to which data owners are exposed by harvesters.

For the record: I am in no way endorsing all of the practices I will document here apart from the legal harvesting of publicly available information under explicit re-use rights and the direct confirmation of all such data. This white paper is simply a description of the many practices I have encountered over the past 15 years and some examples of what I believe are the most effective approaches for those of us in the information industry.

Benefits

The benefit of using harvesting to gather new data or to append and/or confirm existing data is that it decreases the amount of time required to do traditional data acquisition/append/update work. There is no doubt that there is a wealth of information available on the web and, as the amount of information on the web has increased, the potential benefit in effectively harvesting this data has also increased. Specifically, in terms of corporate and biographic research, the increase in the number of companies with web sites, people with micro-sites (LinkedIn, Facebook, blogs), government online services, and news sources able to post content for “free” (i.e., on an ad-supported basis) has made the amount of information potentially “available” exponentially greater than even a few years ago.

The success of a specific harvesting effort, however, is dependent upon several critical factors:

- o What sources are you harvesting?
 - The Internet in general
 - Specific corporate URLs
 - Aggregated news archives
 - Structured online directories
 - Primary source documents in Word, PDF, or other formats
- o What data are you passing to the source to harvest?
 - Company names
 - People's names
 - Industry or geographic parameters
 - Format parameters
- o What data are you getting back from the harvesting effort?
 - Are you getting the records you want?
 - Are you getting the fields of data you want?
 - Is the data you get consistent and reliable or is it contradictory and duplicative?
 - Does the format of the returned data require conversion?

If you are like most harvesters you will soon experience the fact you can easily get consistently formatted and fielded data, but new records may not be in the "universe" desired, the content to added, appended, or updated is very likely to be old, and there may be copyright issues in using the data. And, if the data is in the right universe, "fresh," and without usage issues, it is likely to have major fielding and formatting challenges or come from sources that are difficult to "mine."

For these reasons, the folks who are benefitting from harvesting now are those in businesses where quantity trumps quality and data can be "good enough," but not publication-quality. These include:

- Spammers grabbing email addresses for shot-gun blasts to millions of recipients.
- Other quantity-driven data resellers and online publishers with business models based on displaying individual or groups of records online in exchange for Google AdWords revenue and bulk access to semi-accurate data.
 - Those serving recruitment researchers (e.g., ZoomInfo or the many company-specific efforts that mine job postings, message boards, and classified advertising boards)
 - Those aggregating product information or news sources/blogs by categories

These players do not need to map harvested data to existing records, which can be tricky, and they can actually download absolutely everything from a particular source and pass it along to their visitors/buyers “as is.” Harvesters then typically do mass email efforts to yield a handful of high-value leads or put up enough pages on the web to attract enough eyeballs to make their Google AdWords model profitable.

Traditional publishers are not among those reaping major benefits from harvesting right now because they:

1. usually put up smaller quantities of reliably accurate information under one of two business models (free access to content for qualified, often registered, viewers; or paid access to the content) rather than relying on AdWords revenue
2. can't risk mining competitive sources for legal reasons, and
3. can't afford for their businesses to be associated with spamming

However, there are still major benefits in traditional publishers doing safe, accurate, legal, harvesting as a way to add depth to their records and decrease their overall research costs.

Gathering New Records: This is perhaps the best use of data harvesting, either from the Internet in general or public records in electronic format. In this case the data is gathered via an automated mechanism and then the traditional research process begins. Since the initial harvesting cost is far less than purchasing a list with re-use rights or purchasing a list and mailing a questionnaire, there is an immediate cost savings. Of course, the use of a really inaccurate source would inflate the expense of a traditional research effort (which should involve an initial qualification review, Internet research, telephone research, and quality assurance), but otherwise it is a sound way to gather new records.

Also, it is important to note the large difference between gathering records in a completely new area (new industry segment, new geographical area) where you have no pre-existing records, which is easy, and the cost associated with expanding the penetration in an existing area (e.g., increasing coverage from 70% to 95% for all companies in a given segment). The former is relatively easy and the latter gets progressively more difficult as the penetration level approaches 100%. In the end, it is manual research (Internet and telephonic), preferably by staff with domain experience, that is the only way to get close to 100% coverage.

Data Appending: Publishers typically have highly accurate databases and are constantly trying to append additional data to those records. Typical appends are telephone numbers, emails, URLs, branch offices, business descriptions, and contact names. This strong base helps the harvesting effort because in order to get accurate information you need to start by passing a strong piece of

information to the sources to be harvested. This can be a full legal company name, a telephone number, a personal name, or a URL.

Particularly effective strategies for finding information potentially worth appending are:

- Harvesting contact names and other information from a known URL by searching for text adjacent to certain text strings on the spidered pages (“contact us”; “email”; “about us”; “@”);
- Searching the Internet for certain values in the <title> tag of a page (legal company name, full personal name);
- Searching the Internet by telephone numbers and/or “fuzzy” name (not the exact legal name but all similar variants); and,
- Searching structured online directories by street address or multiple parameters (exact zip code, first x digits of company name).

Once matches are found they need to be resolved via manual research.

The use of “keys” is also very effective in doing matching between databases. This involves harvesting an entire source, fielding the data the same way as the primary source, creating text string “keys” using pieces of the fields (e.g., first x digits of name plus zip code) and then comparing the keys to find matches.

Data Updating: Using harvesting for updating records is accomplished in much the same way as appending. Two additional related tools are the use of URL monitoring and news monitoring to alert researchers to potentially changed data. These mechanisms can be very effective and involve setting up alert services and having resources field incoming changes every day.

At this point it is probably worth stating clearly and in no uncertain terms that harvesting definitely CANNOT automatically update or append a record. This strangely common fallacy is based on a bunch of incorrect assumptions and a lot of wishful thinking.

Risks:

The risks in harvesting are two-fold: the risk to your company in doing harvesting; the risk of being a victim of harvesting.

Harvesting Risks: The primary corporate risks are legal and potential harm to one’s reputation (with one’s competitors if you appropriate their data or one’s customers if you add improperly vetted data). If the data being harvested is “publicly available” for use and you gather it in a proper way, then there is no legal risk. Of course the definition of what “available” means in the context of business usage is an incredibly important one.

Here is my understanding of what is common practice in the US among both the scrupulous and not-so-scrupulous harvesters:

- **Text/Data:**
 - A one-time posting of a subset of information (not a complete copyrighted article or document) can be posted on another site (under free, registered, or paid access terms) with a citation to the source. The content used does not have to be publicly viewable on the web (i.e., it can be accessible via free, registered, or paid access). This falls under the standard US copyright definition of “fair use” and is clearly acceptable. Blogging has made multiple postings from “premium” sources more and more acceptable, especially when these are couched in commentary, linked to the primary source, and/or organized in proprietary taxonomies or included with some other “added value.”
 - Use of any address information for an organization – under free, registered, or paid access terms – is clearly defined as acceptable under the *Feist* decision *unless* an entire source is copied using the same definition of records that are included (e.g., industry, geographic, size, “type” criteria), the fields of data included for each record are the same, *and* the data has not been confirmed via a direct means. Using the same exact definition of records, but excluding the street address, telephone, email and URL would, therefore, still be a violation, although it might not lead to a lawsuit from the original content owner. The “act” of harvesting, however, almost always violates the “terms of use” agreements of web sites because it expropriates server resources without any benefit to the content owner. The *Feist* decision involved keying in or using OCR technology on a *print* product, so there is an important distinction there that would put a harvester from a competitive firm in a tenuous legal position if caught in the act of harvesting. For this reason, this type of harvesting is often done by foreign subcontractors and these subcontractors often use IP-masking and other means to make it difficult to trace who actually did the harvesting. The use of “seed” records (clearly unique records that are specifically created by a content owner for tracking purposes) is common practice in the list rental business and remains the publisher’s best way to track the ultimate beneficiaries of harvesting and to prevent misappropriation by competitors. Another increasingly common and amazingly effective practice is to deliver completely incorrect data when automated requests are detected. (Automated requests are often recognizable by their speed, IP addresses, and methodologies.)
 - The use of data for “inferential” purposes is another practice, although it’s relatively rare. This involves harvesting data from

multiple online directories, mapping records to each other (via specific fields or keys), doing manual research to determine the accuracy of the mapping and the harvested data, and then inferring information about the mapped records and their relationships to the primary record. In other words, if four sources are harvested and then successfully mapped to the primary records and it is determined that when three of the four harvested sources have a different street address than the primary source the primary source is likely to be incorrect then it can be inferred that in all of these cases the addresses should be called for confirmation before other records. This violates terms of use agreements at the source if non-public sources are used, but is a successful (if arduous) way to prioritize records for updating.

- The use of harvested email addresses as a means of “jump-starting” an opt-in email list has been common practice for some time and involves the sending of a “negative option” email to a recipient who has not specifically requested this contact. Basically it is a one-time “spam” effort that informs the recipient that they need to take an action to prevent future emails (as opposed to opting *in* to receive future emails). Those who opt out are removed immediately and other recipients always have the opportunity to opt-out at any point in the future. A corollary is the emailing of people/companies listed in a directory without an opt-in process where the emails are gathered through a manual research process or the publication and sale of these names. These are all violations of the Can-Spam rules, but fall into a bit of a grey area since the recipient is usually “qualified” (chosen for specific reasons and likely to be receptive), they often receive a benefit (a free listing that markets their firm; free industry news), they often post their email addresses publicly, they often have some sort of relationship with the sender (have contacted them; are listed in their publication), they can easily opt-out or block the sender, and the sender is a known firm that can be contacted.
- Images:
 - If posted with explicit re-use rights like Creative Commons then re-usable in any context and often alterable.
 - Publicity photos for a corporate executive or publicly/officially posted image of prominent public figure or the logo of a corporation or public institution are fair game when they are unaltered apart from size (when dimensions/ratios are kept the same). While a harvester who re-used the image would need to suppress any image whose owner objected to the re-use, it is

unlikely to lead to a lawsuit given the publicity intent of the image owner and the inability to demonstrate damages from the re-use.

- If an image is posted by the owner via a self-updating mechanism this is fine, although it requires a manual approval process to prevent the posting of inappropriate materials.

Misappropriation Risks: The most obvious risk is that of a harvester reselling your data and therefore decreasing your revenues. This is relatively rare due to the risk to resellers and their potential buyers, although the spamming business has been generally immune to prosecution.

Conclusion

There are very real benefits to legal, effective harvesting methodologies used in conjunction with traditional data acquisition and data hygiene practices. The amount of time spent on records can be drastically decreased by automated rather than solely manual research.

Below are two examples of the results from a simple automated Internet search using a company name:

EXAMPLE 1

<p>Pantheon Airways</p>	<p>15 Valaoritou 10671 Athens Greece Phone : +30 21 03630123 Fax : +30 21 03630685 Legal FormSociete Anonyme Issued Share Capital60 000 000 EUR VAT number:998939913</p>
<p>Pantheon Airways</p>	<p>Pantheon Airways prepares for takeoff The government is preparing to reactivate the Pantheon Airways scheme to provide a successor to Olympic Airlines, according to Transport Ministry sources. The plan to be submitted to the European Commission provides for the concession of the operations and rights of Olympic Airlines to Pantheon. The size of the new venture will be roughly 50 percent of today's Olympic, against a Commission requirement of 40 percent. The Olympic group will also conduct a voluntary exit program to reduce its staff by about 2,300 employees (2,000 administrative and 300 crew). It is not yet clear how the program will be funded. Speaking to Olympic's pilots last Friday, Transport Minister Costis Hatzidakis said the plan would be tabled to the Commission in Brussels by the end of the month. Hatzidakis also estimated that the activation of Pantheon will begin in October and for a period of six months the two companies will operate in parallel. Against this background, Economy and Finance Minister Giorgos Alogoskoufis last Friday tabled an amendment in Parliament which will protect Pantheon against any claims from Olympic's creditors.</p>

Hatzidakis further told the pilots that some 100-110 of them could be dismissed and requested their union's view about the terms for the pilots' exit.
The union asked the minister for a specific proposal so that they can discuss it at a later date.

This example shows how a researcher can quickly see several potential valuable pieces of data that can then be directly verified by telephone without having to spend many minutes doing manual Internet research.

Example 2 below shows the results of another time-saving automated Internet company name search specifically designed to retrieve URLs to be appended to the records of small and mid-sized UK firms.

EXAMPLE 2

A & D HOPE (SCS.) LTD	http://www.adhope.com
A & D JOINERY LTD	http://www.aanddjoinery.co.uk/
A & F GRANT LTD	http://www.afgrant.co.uk/
A & J GUMMERS LTD	http://www.sirrusshowers.com/
A & J SCIENTIFIC LTD	http://www.dsr.co.uk/
A & R VEHICLE SERVICES LTD	http://www.aandrvehicleservices.co.uk/
A & S PACKING (YORKSHIRE) LTD	http://www.aspacking.co.uk

Searches of specific corporate URLs are extremely useful for extracting email addresses and contact names. In this next example the names extracted from the corporate URL and the specific URL within the corporate site is cited.

EXAMPLE 3

Lami Wood Products Corp.	Robert Alexander, President; Bob Alexander, VP; Bill Luca, VP of Sales; Patrick Whitchurch, Director; Dan Heath, Controller; Kathy Ferrell, Corporate Secretary; Michael Stuecken, IT Manager; Jerry Shirrell, Manager; Jeff Alexander, Manager .	http://www.signaturekb.com/lamiloc.html
AST/Adhesive Systems Technology Corp.	James J. O'Brien, Chairman and Chief Executive Officer; Lamar M. Chambers Sr. Vice President and Chief Financial Officer; David L. Hausrath, Sr. Vice President and General Counsel; Robert M. Craycraft, Vice President and President, Ashland Distribution; Susan B. Esler , Vice President, Human Resources and Communications; Samuel J. Mitchell Jr., Vice President and President, Ashland Consumer Markets	http://www.ast-corp.net/profile.html
Winona ORC Industries, Inc.	Mary Malloy, Chairman; Ron Wenzel, VC; Jim Pomeroy, Treasurer; Chad Anderson, Manager, Mike Kreiling, Manager; Jim Yenish, Voting Member; Kevin O'Reilly, Voting Member; Blaine Krogh, Advisor Finance; Richard Enochs, Voting Member; bharris@worcind.org	http://www.worcind.org/html/our_location.html

Publicly available government sources are rich sources of timely, accurate, and valuable information. Here is an example of a search of a government database (daily Federal Aviation Administration ownership transfer reports in PDF format) to extract specific fields of data. In this case a 2000pp document can be processed in under 1 hour to extract extremely valuable and timely data using automated processes and manual review.

EXAMPLE 4

<i>FAA Registry #</i>	<i>Price Paid</i>
N464E	\$340,000
N30251	\$100,000
N3656Z	\$10,500
N30PX	\$75,000
N5470K	\$44,000
N97931	\$12,000

These are just a few of the many effective and legal ways that harvesting can help information businesses to decrease their data maintenance and acquisition costs while they increase the value of their content to their end-users.

--

Matt Manning is the president of Information Evolution, Inc., a firm that designs and implements efficient research and editorial processes for content companies.