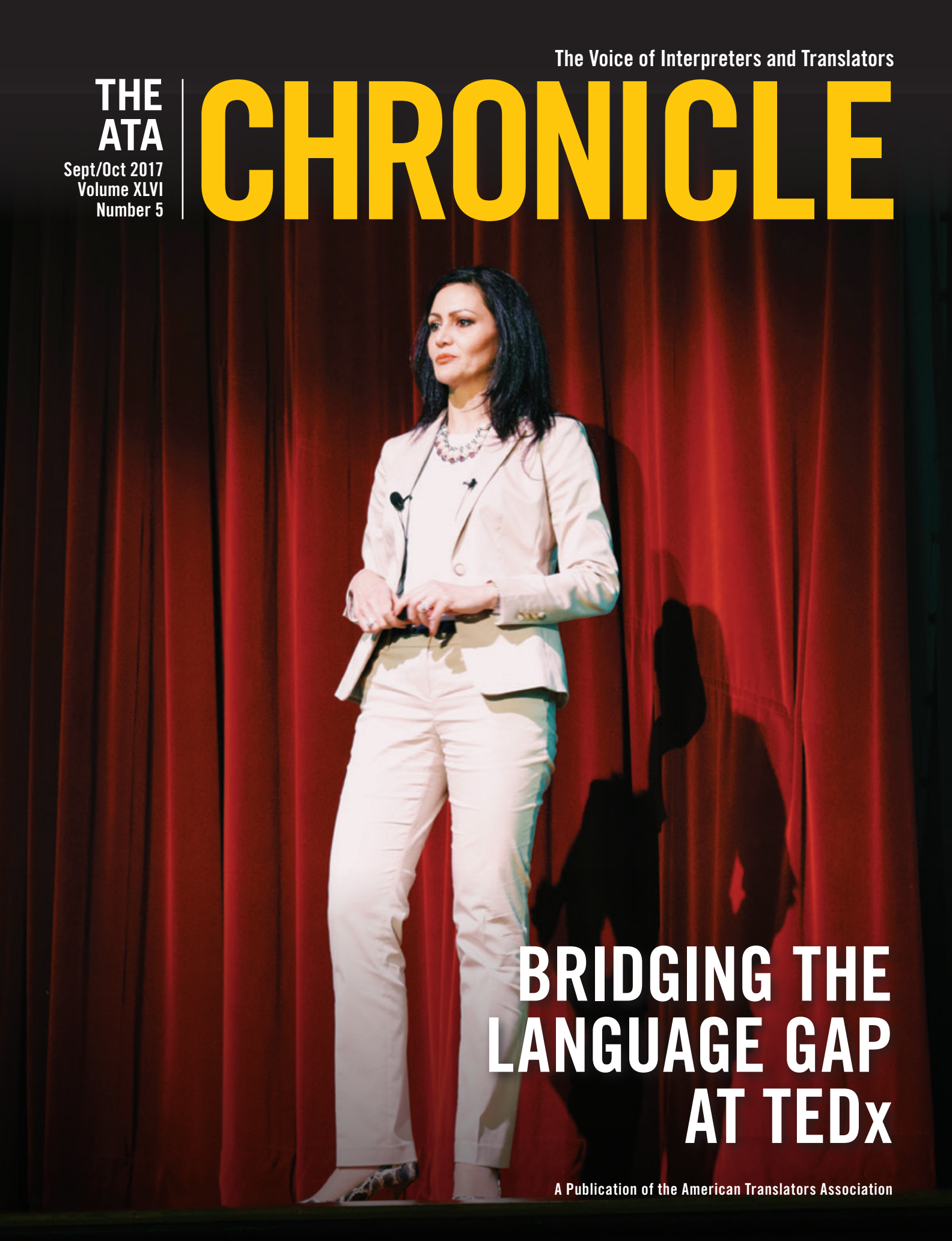
The Voice of Interpreters and Translators

THE ATA

Sept/Oct 2017
Volume XLVI
Number 5

CHRONICLE



BRIDGING THE LANGUAGE GAP AT TEDx

A Publication of the American Translators Association

Who Is Really Visiting Your Website? (It's Not Who You Think!)

Learn how Google Analytics can mislead you if you're not careful.

Over the past decade, several articles in *The ATA Chronicle* and sessions at ATA Annual Conferences have addressed the importance of websites as a way to market translation and interpreting services.[1] As websites have become more important to translators and language services providers, many of us have tried to measure how successful our websites are at attracting new business.

A variety of monitoring and analytical tools are available for this purpose, but one of the most widely used is Google Analytics. Unfortunately, within the past two or three years unfiltered Google Analytics reports have become increasingly contaminated by automated computer programs called "bots" (also known as "spiders" or "crawlers").

What is a bot? Bots are software applications that are written to perform specific online tasks, usually repetitive ones that would be impossible or difficult for humans to perform quickly. While "good" bots perform useful functions similar to those initiated by search engines to index a website, "bad" bots visit websites with all sorts of evil intentions for disrupting internet traffic, such as spamming, content scraping, and malware distribution.[2] This can cause problems in terms of analytics, particularly in reports measuring who is visiting your site. These contaminated reports can lead to potentially bad marketing decisions for the companies that take the results at face value. As a result, the accuracy of Google Analytics reports could be dubious. This was the case at our company. But before relaying our story, let's back up a bit and explore a few basics about search engine optimization.

EARLY ATTEMPTS TO DRIVE TRAFFIC TO WEBSITES

About 15 years ago, website owners attempted several techniques to drive traffic to their sites. Frank Dietz, in his seminal article in *The ATA Chronicle* in 2006 ("Search Engine Optimization for Translators and Interpreters"), identified a number of these techniques, including:

- Using meta tags (snippets of text) to describe the content of each page.
- Identifying key words and using them in expanded website text.
- Registering with multiple search engines, such as Google, Yahoo, MSN Search, and AOL Search.
- Creating inbound and outbound links to the sites.
- Improving internal site navigation.

Many of the techniques mentioned by Dietz are still relevant today.[3]

The activities Dietz describes culminated in search engine optimization (SEO), which is "the process of affecting the visibility of a website in a search engine's unpaid results," often referred to as organic (search) results.[4] Basically, SEO means getting your site to appear as one of the first suggested sites when someone searches with Google or another search engine. For several years users were bombarded with marketing propaganda from the purveyors of SEO, promising improvements in keyword rankings, link popularity, organic traffic, and even visibility on the first page of Google, Yahoo, and Bing.

How does SEO work? SEO projects typically begin with a keyword search that involves the identification of somewhat odd phrases that people actually enter into search browsers (e.g., "online English-to-Spanish document translation"). The next step is to modify existing website text to include the odd phrases. This is followed by the generation of "new" content (press releases, articles, blogs, etc.) with links to the "optimized" website that are then placed on a variety of sites, including those that are

frequently visited, as well as obscure sites that will publish almost anything.

With the exponential increase in website and related content, plus billions of links, search engines use ever-changing search algorithms to increase the probability that users receive quality links to websites with "killer content." Since each change in a search algorithm alters search results, SEO is never finished.

Once the effects of "bad" bots and spam referral data are removed, you'll have a much better idea about the origin of visitors to your website and actionable data.

MEASUREMENT OF WEBSITE TRAFFIC

Attempts to drive traffic to websites have been accompanied by efforts to measure the success of such attempts. At first, online services provided users with one-dimensional data, such as website hits and the number/source of inlinks. (An inlink is a link directed to a website. The more inlinks any given website has, the more likely that the website will rank higher in a search ranking.) The launch of Google Analytics, a service that tracks and reports website traffic, in November 2005 shook the market. Within one week of its launch, Google Analytics signed up 100,000 new accounts, which was four times larger than the entire website statistics market. Edging out competitors such as WebTrends, Coremetrics, Omniture, and IBM, Google Analytics quickly monopolized the website statistics market with its free services. By 2015, 30 million websites were using Google Analytics.[5]

EXPERIENCE WITH GOOGLE ANALYTICS: HAD IT AND LOST IT!

Our company relied on the website statistics generated by Google Analytics for several years. We found that the reports it generated helped us understand the success (or failure) of attempts to drive traffic to Inline's site.

However, as mentioned earlier, the initial usefulness of Google Analytics reports to our company has declined significantly. This deterioration began in 2014 and accelerated in 2015 and 2016. How did we determine this? The remainder of this article will offer a glimpse into how we uncovered contamination in the reports generated by Google Analytics with our company's actual data and how we worked to solve the issue. By detailing our story, we hope to leave readers with some pointers on how to recognize contaminated data and how it makes it difficult, if not impossible, to assess the performance of your website. We'll also offer solutions that freelance translators and smaller translation companies can implement.

FIRST ATTEMPTS TO DRIVE ONLINE TRAFFIC

From the start of 2009 to the end of 2012, our company successfully used Google Analytics reports and data to monitor the effectiveness of attempts to drive traffic to our site. (See Figure 1 below, which shows the number of site visits during this time period.)

Our company's first attempt to increase website traffic began in late 2008, when we contracted with DirectoryM, a specialized advertising company that emphasized online marketing campaigns to businesses. DirectoryM's affiliation with online business journals across the U.S., coupled with small ads with links to our company's website, resulted in an initial uptick in visits to our website during a one-year trial.

The second major attempt to drive traffic to our company's website began in early 2010. We could no longer resist the siren call of the SEO consultants with their promises of hordes of new customers descending upon a highly visible website. So, we engaged a SEO consulting firm for a one-year program that included the usual services: keyword research, website modification, and content generation. Although the SEO consulting firm sent us a stream of reports showing how successful they had been in their efforts to increase traffic to our site, we continued to use Google Analytics reports to independently measure the impact of our newly optimized website. After the one-year SEO program, which cost about $30,000 and resulted in only $590 of new business, we decided not to renew the contract.[6]

GOOGLE ANALYTICS: THE GOOD YEARS

In addition to site visit statistics, Google Analytics reports have provided our company with detailed visitor profiles, including the languages spoken by site visitors, how long visitors stayed on the site, the source of traffic, new versus returning users, the pages visited and the sequence of pages visited, the percentage of single page visits (called the "bounce rate"), and much more. Of all the statistics provided, the Source/Medium report, which shows the origin of site visits and referrals, was initially the most useful and actionable.[7] The 2009 and 2011 data

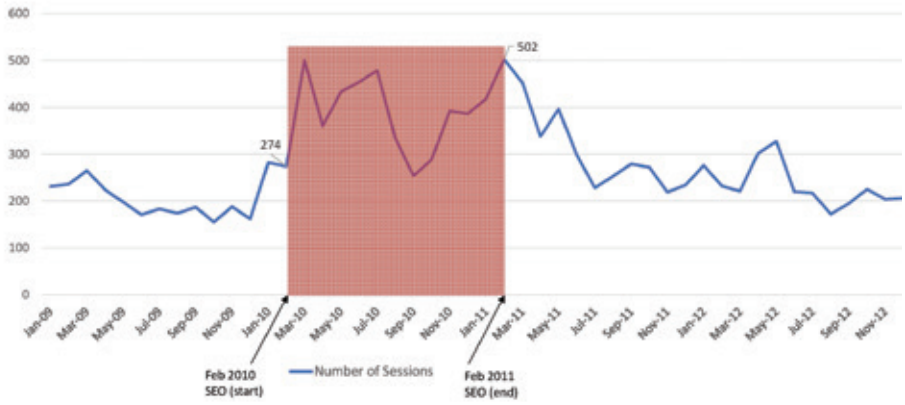


Figure 1: Number of site visits by month before, after, and during SEO program (2009–2012) as reported by Google Analytics

columns in Figure 2 on page 22 contain easily recognized websites, such as ATA (atanet.org), proz.com, translatorscafe.com, and linkedin.com. In addition, organizations we cultivated for referral links for new business or for recruiting are also present. For example, a regional alliance of family-owned print shops (cprintalliance.com) appears in the 2009 data, and the Middlebury Institute of International Studies at Monterey (miis.edu) appears in the 2011 data. The term "organic" means that the site visitor found our site as a result of a particular search strategy that they initiated when using Google, Yahoo, or Bing. The term "direct" simply means that the visitor came directly to our site by entering our company's URL in their browser.

THE HIJACKING/CONTAMINATION OF GOOGLE ANALYTICS

Beginning in 2014, bots and spam referrals began to distort our Google Analytics data. Apart from "organic searches" via Google and "direct" site visits, we couldn't recognize what was really behind the main sources of visits to our company's website in 2016. Historically important sources of visits, such as atanet.org, translatorscafe.com, and linkedin.com, were no longer in the top 10 and seemed to drop in importance. What we didn't realize at the time was that they were displaced by bad bots that had targeted our site (and sometimes Google Analytics separately) with hundreds of visits of extremely short duration, leaving behind a trail of what are called spam referrals. In fact, eight of the top 10 referral sources in 2016 consisted of bad bots, such as rank-checker.online, site-auditor.online, monetizationking.net, and traffic2cash.xyz. (See the 2016 data column in Figure 2.)

Prior to our research for this article and discovery of just how contaminated Google Analytics data can be, we were seriously considering measures such as consolidating our three ATA memberships into a single membership based on the decline in ranking of atanet.org. Fortunately, general inertia prevented us from making such an unwise decision.

THE RUSSIANS ARE COMING... OR ARE THEY?

In addition, the profile of visitors to our site changed from overwhelmingly English-speaking in 2009 and 2011, to majority Russian-speaking, according to Google Analytics reports. (See Figure 3 on page 23.)

The surge in Russian speakers to our website from 2011 to 2016 bore no relationship to the types of languages requested by our clients. English>Russian and Russian>English translation projects have accounted for less than 1% of our business from 2009 to the present.

The increase in bot traffic is highly correlated with the surge in Russian-speaking visitors. (A smaller surge in Brazilian Portuguese visitors can also be seen for this same period in Figure 3, which is also unrelated to any increase in Portuguese translation work.) Although correlation is not a guarantee of causality, it may be possible to dig deeper into Google Analytics to connect these two events.

RECOGNIZING BAD BOTS AND SPAM REFERRALS

Bad bots and spam referrals are relatively easy to identify. Most will have unfamiliar names, often containing telltale words such as "rankings," "traffic," and "cash." Indications of bad bot traffic in Google Analytics reports include low session durations, high bounce rates (i.e., visitor leaves after visiting only one page), an unexplained surge in new visitors with low engagement (combination of high bounce rate, short session duration, and no goal completion, such as viewing a certain number of pages per visit or going to a specific URL), or use of an xyz domain.

Figure 4 on page 23 shows how the profiles of sites with real referrals differ from sites spewing out spam referrals. Note how referrals from legitimate sites, such as atanet.org and cprintalliance.com, lead to sessions that last more than a minute, with visits to multiple pages, and submittal of an occasional request for a

2009		2011		2016	
Source/Medium	Sessions	Source/Medium	Sessions	Source/Medium	Sessions
1. google/organic	1,063	1. google/organic	1,502	1. google/organic	1,161
2. direct	558	2. direct	863	2. rank-checker.online/referral	800
3. atanet.org/referral	94	3. bing/organic	328	3. site-auditor. Online/referral	791
4. bizjournalsdirectory.com/referral	81	4. yahoo/organic	267	4. direct	566
5. articles.directorym.com/referral	56	5. atanet.org	71	5. monetizationking.net/referral	387
6. yahoo/organic	53	6. miis.edu/referral	60	6. traffic2cash.xyz/referral	84
7. cprintalliance/referral	49	7. translatorscafe.com/referral	53	7. keywords-monitoring-your-success.com/referral	80
8. proz.com/referral	48	8. search/organic	48	8. website-analyzer.info/referral	73
9. bing/organic	34	9. linkedin.com/referral	45	9. uptime.com/referral	61
10. directorym.net/referral	27	10. proz.com/referral	39	10. fix-website-errors.com/referral	57
				13. atanet.org/referral	51
Average Session Duration: 112 seconds		**Average Session Duration:** 56 seconds		**Average Session Duration:** 40 seconds	

Figure 2: Source/Medium Report (origin of site visits) and average session duration for 2009, 2011, and 2016 (unfiltered Google Analytics data)

quote (i.e., one of the main goals of having a website in the first place!). In contrast, referrals from bad bots, such as rank-checker.online and website-analyzer.info, lead to sessions lasting just a few seconds, with most visits to only one page, and never result in a quote request.

HOW TO FILTER OUT CONTAMINATED INFORMATION FROM GOOGLE ANALYTICS REPORTS

Unfortunately, there is no universal fix or global solution for eliminating bad bots and spam referral from Google Analytics data. Nevertheless, we provide three ways you can "decontaminate" Google Analytics data. They range in cost and complexity and the choice will ultimately depend on your objectives and the resources you want to devote to the effort.

Apply Filters: The more advanced Google Analytics users and aspiring cyber security sleuths can visit the ADMIN section of Google Analytics. (See Figure 5 on page 24.) Under the "View" column, you will find a section called "Filters." Here, filters can be added to remove additional unwanted bot traffic and spam referrals. At this point, the steps involved can be stressful and time-consuming and could have unintended results such as accidently removing a legitimate website crawler that is indexing your site and which could increase its visibility. Consequently, you should consider taking advantage of Google's automatic exclusion feature discussed in the next paragraph, or contracting with experts on the matter. Keep in mind that the creation of these filters will not clean up historical data.

Check the Box: In July 2014, Google Analytics announced a new feature to automatically exclude bots and spiders that appear on what is known as the International IAB/ABC Spiders and Bots List[8] that is updated continuously. Since every Google Analytics account requires a unique sign in, there is no direct link to the box you need to check to exclude these nasty critters. You can find this feature by going to the "ADMIN" section of analytics and clicking on "View Settings" under the "View" column. (See Figure 5 on page 24.) Under "Basic Settings" you can check the Bot Filtering box labeled "Exclude traffic from known bots and spiders." Unfortunately, "Checking the box" will not clean up historical data, but should filter data going forward. It may take a week or two to begin working. You may want your website administrator to set up a test view first to prevent any accidental damage to unfiltered historical data you may need later. Note: You will need admin access to your Google Analytics account to apply filters or to check the Bot Filtering box.

2009		2011		2016	
Language	**Sessions**	**Language**	**Sessions**	**Language**	**Sessions**
1. English (U.S.)	1,819	1. English (U.S.)	3,283	1. Russian	1,643
2. English	57	2. English	105	2. English (U.S.)	1,417
3. Chinese	49	3. English (U.K.)	54	3. (not set)	817
4. German	43	4. Spanish	53	4. Russian (Russia)	474
5. Spanish	43	5. Chinese	51	5. Brazilian Portuguese	173
6. Italian	39	6. Spanish (Spain)	47	6. Secret.google.com You are invited! Enter only with this ticket URL. Copy it. Vote for Trump!	114
7. Brazilian Portuguese	37	7. Brazilian Portuguese	34	7. Chinese	57
8. French	34	8. French	30	8. c *(the "c" indicates a bot)*	55
9. Spanish (Spain)	31	9. German	23	9. Spanish	52
10. Russian	31	10. Russian	23	10. English (U.K.)	37
Average Session Duration: 112 seconds		**Average Session Duration:** 56 seconds		**Average Session Duration:** 40 seconds	

Figure 3: Audience Overview: Language–2009, 2011, and 2016 (unfiltered data)

Source	# of sessions	Average # of pages per session	Average session duration (minutes:seconds)	# of quotes requested
atanet.org	553	2.99	1:38	7
cprintalliance.com	182	2.13	2:40	17
rank-checker.online	800	1.01	0:12	0
site-auditor.online	791	1.18	0:17	0

Figure 4: Contrasting profiles of visitors from legitimate websites and profiles of the bad bots that generate spam referrals (2009 to 2016)

Manually Recreate the Source/Medium Report: Since you cannot apply filters to the contaminated historical data, you may want to recreate the Source/Medium report for one or more past periods. This can be done by opening a new document in Word and creating a simple table. Copy the valid sources of site visits and data from the contaminated Source/Medium report, while purging data left behind by the bad bots and spam referrals. You can then recalculate session numbers and re-rank the sources. Bad bots and spam referrals can be found by looking for characteristics such as a 100% bounce rate, session durations of one second or less, and the telltale words mentioned earlier in this article. The Source/Medium page may be the only one you can recreate, but it's one of the most important reports that provides actionable information for marketing your translation business.

In Figure 6 on page 24, we compare our company's manually filtered 2016 Source/Medium data with 2016 unfiltered data. Note how atanet.org regains its historic position in the top five sources of referrals, as compared with the 13th position in the unfiltered data.

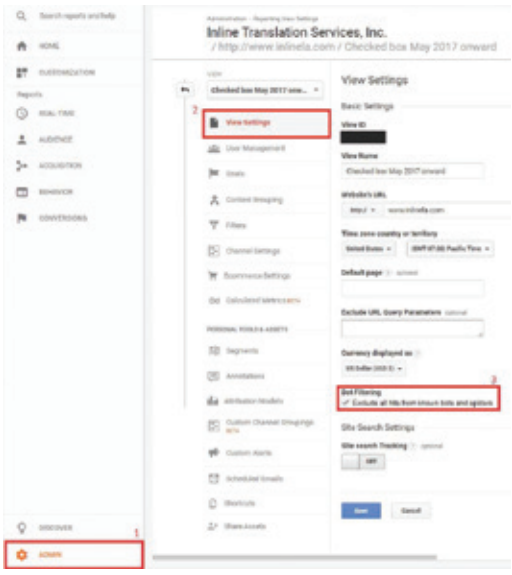


Figure 5: Applying the Google Analytics Bot Filtering Feature

2016 (unfiltered)		2016 (manually filtered)	
Source/Medium	Sessions	Source/Medium	Sessions
1. google/organic	1,161	1. google/organic	1,161
2. rank-checker.online/ref	800	2. direct	566
3. site-auditor/referral	791	3. atanet.org/referral	49
4. direct	566	4. bing/organic	47
5. monetizationking.net/ref	387	5. translatorscafe.com/referral	35
6. traffic2cash.xyz/ref	84	6. linkedin.com	32
7. keywords-monitoring-your-success.com/ref	80	7. proz.com/referral	20
8. website-analyzer.info/ref	73	8. yahoo/organic	14
9. uptime.com/referral	61	9. paymentpractices.net	13
10. fix-website-errors/ref	57	10. yelp.com/referral	13
13. atanet.org/referral	49		

Figure 6: Audience Overview: Source of Visitor in 2016 as reported by Google Analytics (before and after purge of bot traffic and spam referrals)

WHAT'S IT ALL MEAN?

Google Analytics data has increasingly become contaminated by the emergence of bots and spam referrals. This contaminated data creates lots of noise and drowns out important signals from the marketplace. For this reason, Google Analytics reports should not be used in an unfiltered format.

Depending on your needs, you may want to add custom filters to your Google Analytics program, which you can do in the ADMIN section of your Google Analytics account. This approach, however, requires time spent researching bots, identifying the bad ones, and then excluding them with proper coding. Even then, this "do-it-yourself" approach won't prevent your data from becoming contaminated. The bots are evolving constantly so your work will never be finished, even with expert help.

Consequently, we highly recommend you consider taking advantage of the Google Analytics Bot Filtering feature, which you can activate as described above. (The default mode is OFF.) The advantage to this approach is that the list of excluded bots is updated constantly without further intervention on your part.

Even if you don't apply filters to your Google Analytics program, it's possible to manually recreate the Source/Medium report, which is one of the most useful Google Analytics reports. After removing the residue from the "bad" bots and spam referrals, you'll have a much better idea about the origin of visitors to your website and actionable data. And you may not have to learn Russian after all!

NOTES

1 Berrill, Simon. "Spider Marketing: How to Get Clients to Come to You," *The ATA Chronicle* (January-February 2017), 8-11, http://bit.ly/spider-marketing-Berrill.

2 Čandrić, Goran. "How to Exclude Bot Traffic from Your Google Analytics" (February 9, 2017), http://bit.ly/filter-bot-traffic.

3 Dietz, Frank. "Search Engine Optimization for Translators and Interpreters," *The ATA Chronicle* (April 2006), 28-30, http://bit.ly/SEO-Dietz.

4 "Search Engine Optimization, http://bit.ly/SEO-wiki.

5 Clifton, Brian. "Google Analytics Is 10 Years Old: What's Changed?" *The Insights Blog* (November 10, 2015), http://bit.ly/Google-Analytics-Clifton.

6 Paegelow, Richard. "The Search Engine Optimization Witchdoctors: How to Spend $30,000 and get $590 of New Business," Presentation at the 13th ATA Translation Company Division Conference (Orlando, Florida, January 2013).

7 The Source/Medium report shows the origin of a site's traffic (such as Google) or a domain (such as example.com). It also shows the general category of the source called the Medium: for example, organic search (organic), web referral (referral), and cost-per-click paid search (cpc).

8 A merged database of advertising and site-related non-human (robotic) activity impacting websites that's extremely important for the media industry and for ad agencies and advertisers. See: ABC-Audit Bureau of Circulations Ltd (www.abc.org.uk/about-us) and iab.europe (http://bit.ly/iababc-spiders-and-bots-list).

Richard Paegelow is the managing director and senior Romance language editor of Inline Translation Services in Glendale, California. He has an MA from the University of Kansas (majors: Spanish and Latin American area studies), as well as an MBA from Columbia University. Prior to purchasing Inline in 1992, he worked for 15 years in general banking, management training, and business consulting. He is the chair of the International Programs Advisory Board at the University of Kansas. Contact: richard@inlinela.com.

Thea Dery joined Inline Translation Services in May 2017 as an assistant translation services coordinator. A recent graduate of Occidental College in Los Angeles, she majored in Spanish language and culture. Her undergraduate experience included a semester at the Pontificia Universidad Católica de Valparaíso in Chile. She has worked on a variety of translation projects with public health and environmental agencies in the U.S.