ORACLE®
Data Cloud

Grapeshot: Description of Methodology

Printed on 04/11/2019

Table of Contents

Executive Summary

This Description of Methodology (DoM) is a summary of processes employed for the delivery of the Grapeshot products and services (now part of Oracle Data Cloud). It includes a general description of Grapeshot's scientific and technological underpinnings, key implementations, and ways our partners integrate with our systems. It is not a DoM for any other Oracle Data Cloud products and services.

What's Included in this DoM:

- The Technology
- Processes and Use Cases
- Quality Control and Verification

Grapeshot fundamentally is about contextual analysis of textual content on webpages, processed at the massive scale and speeds required by automated advertising technology. It is a webpage textual classification database technology.

Our technology crawls those pages then uses information retrieval processes to determine and extract the central textual content, what we call the "epicenter" of a page. (We exclude all other elements, such as images, video, advertisements and sidebar content, as explained further below.) We then analyze and compare that epicenter content to "keyword segments," sets of carefully compiled words and phrases concerning specific topics, to determine if there is a match, and if so how strong the match is. We use that information to, in a pre-bid automated advertising environment, indicate whether a page should be negatively targeted (removed from consideration) for a given advertising client, or may have contextual relevance and be positively targeted. Our technology works for webpages on desktop and mobile device environments.

To assure that our technologies remain current and of high quality, are used as intended and that we follow industry standards and best practices as they are updated by oversight bodies such as the MRC we have deployed multiple processes as described below. in MRC parlance, Grapeshot conducts its operations at the "property level" but not at what MRC calls the "content level".[1]

---

[1] http://www.mediaratingcouncil.org/MRC%20Ad%20Verification%20Supplement-%20Enhanced%20Content%20Level%20Context%20and%20Brand%20Safety%20(Final).pdf (page 5)

Primary Users and Use Cases

The overarching goal of Grapeshot's technology and processes is to offer advertising buyers (the "buy" side), media sellers (a.k.a. publishers or the "sell" side), and their advertising technology partners a transparent lens around the meaning of the core textual content of published webpages, and to use that understanding to make decisions about the placement of advertising on those pages.

All sides deploy Grapeshot's technology via technology platforms they use, such as DSPs, SSPs, advertising exchanges, ad servers, and measurement and verification services. Grapeshot's technology is used for both desktop and mobile webpages. For every use case, the fundamental technology is the same. Grapeshot is used for categorization and ranking the main textual content on a webpage against keyword segments that indicate contextual relevance.

It is important to note that other important platform functions such as ad serving, detection of ad fraud, identification of invalid traffic (IVT/SIVT), measurement of viewability, measurement of audiences, and other cookie implementations are not handled by Grapeshot or its technology.

Use Cases

There are two primary business uses for Grapeshot's technology:

- **Enhancement of Brand Safety**. Grapeshot improves advertisers' ability to avoid serving their messages adjacent to or embedded within webpage textual content they may find objectionable.

- **Improved Contextual Targeting**. Advertisers and publishers use Grapeshot to find webpages whose core textual content is relevant to advertisers so that pertinent advertising messages may be served on those pages.

A Use-Case Example

As stated earlier, the overarching goal of Grapeshot's technology is to get at core meaning of a webpage's central textual content and evaluate it in the desired context for the specific application by applying the appropriate ranking and levels of importance to the various terms in the text being analyzed. Said another way, Grapeshot's technology is constructed to analyze not just words but also their meanings and relevance in context.

As one illustration, “ball” is a classic example of a word that can have multiple meanings, at least one of them potentially objectionable, while the others are benign or even desirable for certain advertising circumstances. Grapeshot also allows custom gradations and control. Some brands, for example, may not object to having their advertising message on an article in which the word “ball” is used in even its racy sense, or may be fine with ads placed if the named activity is not the main focus.

Grapeshot's Processes

Overview

To do its work, Grapeshot:

1. Crawls hundreds of millions of webpages daily and indexes the core meaning of their text to gain an understanding of the subject matter and categorize it appropriately. Algorithms are used to identify the relative “weight” of all words within the text (e.g., a news story on a webpage). These weighted words are generated as an atomic composition of that document.
2. Separately creates groups of words and phrases known as “keyword segments.” These are themselves sets of words determined to reflect a particular topic.
3. Matches the understanding of the processed text to the keyword segments and provides scores to indicate the degree of match. Grapeshot will make multiple probabilistic matches between a set of such keyword segments and a document.

Advertisers and those serving them use the technology to avoid having their messages appear in undesirable contexts, or to surface relevant pages in which to target advertising.

The Science Behind Grapeshot

Grapeshot’s core technology is based on Information Retrieval (IR) science developed for the last few decades at the Computing Linguistics and Computer Laboratory Departments at Cambridge University. There, Dr. Martin Porter, a co-founder of Grapeshot, focussed on IR within his fields of study. The “Porter Stemmer” algorithm that bears his name stems words back to their root word, or stem. This linguistic tool, written in 1980, is used in major search engines, including those run by Google, IBM and Microsoft.

Dr. Porter developed new ways of not only processing and understanding language on a page but also doing so quickly and with minimal intrusion into the workings of the page. Grapeshot today can be understood as it was described by Dr. Porter: as a search and information technology built on some well-established IR principles that have stood the test of time. Unlike many semantic solutions, which must generate a set of rules to a core system which can then assess and score documents, Grapeshot's two-step process – processing the page then matching against keyword segments -- may be considered more flexible.

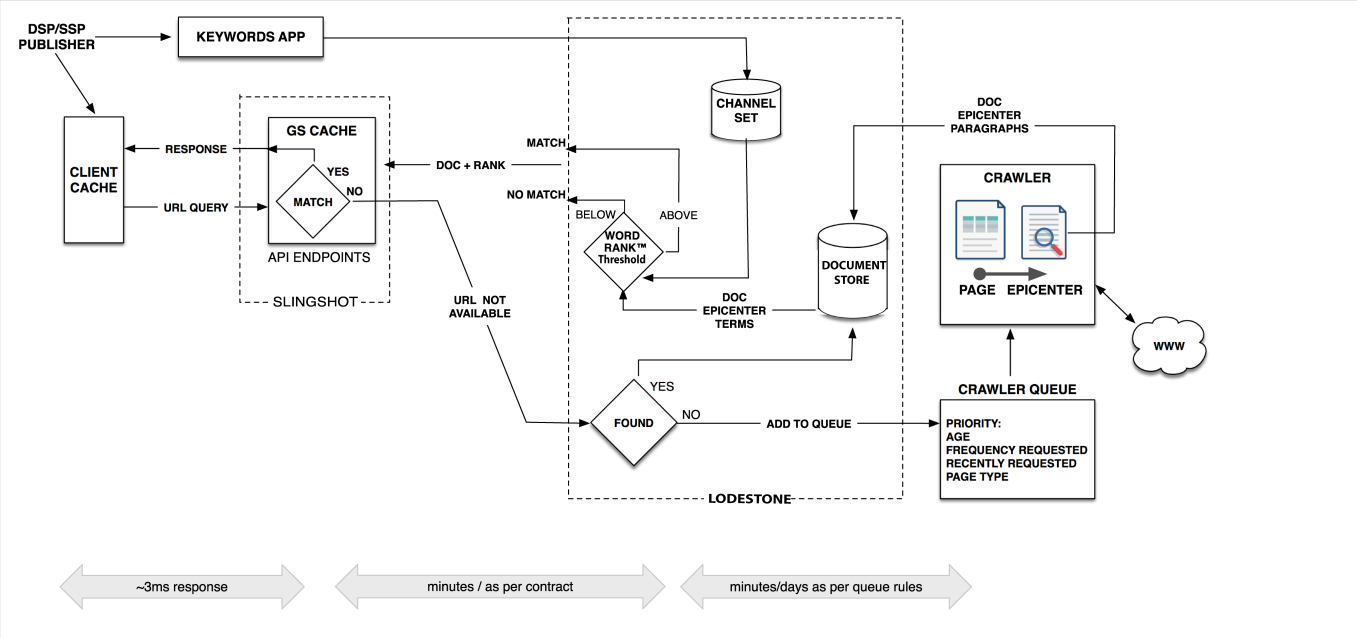
Dr. Porter’s initial mission was to determine the sense, the full meaning of a document, then to apply it to the multi-million query, micro-second, low-latency needs of a real-time bidding (RTB) advertising environment, the fastest and most intense framework in which Grapeshot is deployed.

Grapeshot's Architecture and Flow

The Process

The Layers: Crawler, Categorization and Cache

Grapeshot crawls and contextually analyzes text in response to requests. The technology and architecture support loads exceeding 3 million queries per second (QPS), a level found in massive programmatic advertising implementations, and also offers sub-millisecond response times. Grapeshot delivers this speed and scale through a compartmentalized infrastructure. The architecture subdivides into three basic layers (which we will represent here, looking from right to left):



An overview of Grapeshot's technology.

1. **Crawler layer (right)**: a system that manages the crawling of requested webpages. The system crawls a webpage, and from its HTML extracts the core textual content, also known as the "epicenter." It also manages the optimal order of crawling, and the frequency of re-crawling.
2. **Categorization layer (known as “Lodestone”)**: the core engine that conducts probabilistic matching to keyword segments as described earlier.
3. **Cache layer (known as “Slingshot”)**: a localized installation set up to enable responses for partners at speeds required to deliver responses in real-time-bidding (RTB) advertising environments.

Crawler Layer

The request for a page at a given URL can originate from a number of sources but typically will have been from a partner’s technology platform or server that handles ad serving, measurement and/or verification.*

Pages are crawled at a rate of well over 100,000 per minute, and are re-crawled to check for changes depending on the propensity of a particular page at the given URL to change within its epicenter.

The crawler has a management scheduler, which ensures permission to crawl the page or domain and also ensures the crawler does not overload a page with multiple repeat visits. The crawl information for any individual publisher's website is used for all partner implementations. To account for webpages that are the same as each other but have subtly different URLs

-- such as from parameters toward the end of a URL that are caused by web analytics software -- Grapeshot strips out those parameters.

Sites that Grapeshot is attempting to crawl can exclude or block the crawlers by various means, such as via their robots.txt page or by specifically excluding Grapeshot's crawler. If Grapeshot is unable to crawl a page, information about that inability will be indicated to Grapeshot's partner or customer. (There is further information below on cases in which Grapeshot has found a publisher attempting to thwart or deceive crawling for nefarious purposes.)

Epicenter, Adjacent and Non-Text Content

Once Grapeshot's crawler has received a request to scan an HTML webpage, it finds and crawls that page then downloads the page's core textual content (the epicenter, as described earlier) from the page's HTML. We do not download or analyze the CSS, JavaScript, images, navigation, footer, sidebars, and other areas tangential to the main textual content on the page. (Picture a typical news webpage. Grapeshot's technology will download and analyze the central text of that page but not the embedded or side elements which may include "Related Stories," additional linked headlines, images, videos, and so on.) Said another way, we conduct what the MRC calls "property-level reporting" without the level of measurement granularity that would include code or objects, including those from third parties, or that appear outside, adjacent to, or embedded within the main text on a page.

We inform partners that we analyze a webpage's epicenter, not the surrounding material: the main textual content but nothing else -- not images, videos, graphics, sidebar content, or third-party insertions such as paid advertising. That said, we are aware that surrounding, adjacent, or embedded content on a webpage, content that may be provided by JavaScript executions, or non-textual content such as images or video can be seen to affect the context of a page as presented to users and therefore be a consideration for advertisers.

Prioritizing Crawl Requests

When requests to crawl a page are received they are put into a priority queue, determined by a number of factors:

1. Bookmarklet request. Grapeshot releases Javascript bookmarklets that enable users to, from a web browser, manually initiate a scan of a specific publisher's page for matches to designated Grapeshot keyword segments.
2. Requests through our architecture as described above and below in the "Integrating With Grapeshot" section.
3. Scans of pages whose TTL has expired.

Time to Live (TTL)

Pages change, of course, and need to be re-crawled. The crawler maintains an estimate of how frequently a page changes. If a page has been modified since the last time it was crawled, then the crawling frequency is halved, to a lower limit of every four hours. If it hasn't been modified, then the crawling frequency is doubled, to a maximum of every 30 days. In this way, the rate of re-crawl soon matches the modification rate, providing efficiency in apportioning resources.

Grapeshot conducts empirical tests to assure that the four-hour minimum threshold is sufficient to detect epicenter changes at a rate of 95% or above. Should the tests find meaningful changes on more than 5% of pages, we shall apportion more resources in order to reduce the minimum scan time, also known as the time-to-live (TTL) of our categorization results.

Categorization Layer ("Lodestone")

Once a webpage has been crawled, its information is sent to Grapeshot's categorization layer (which we call "Lodestone"), where the page's information is kept in a document store or corpus -- a central data store of the information from all crawled pages. Grapeshot's central data store holds information from more than 5 billion documents at a time, and is an ever-growing, frequently updated record of all pages that have been crawled.

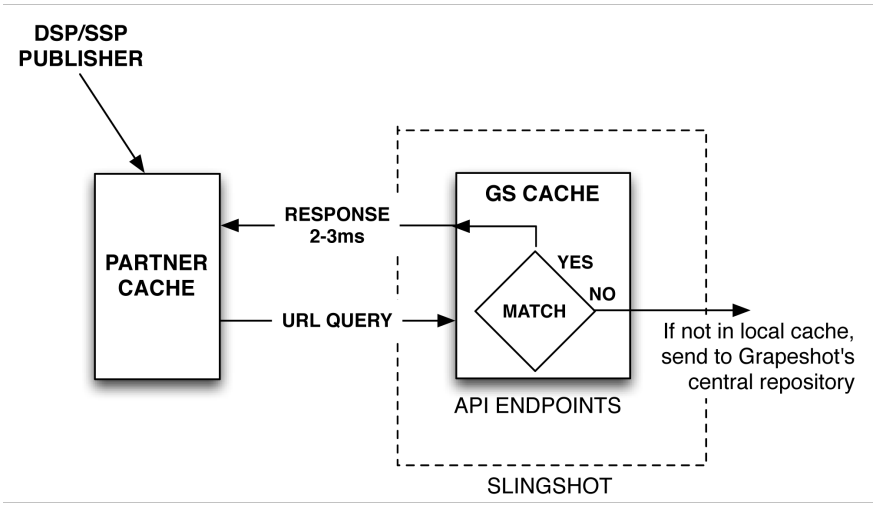
From this corpus, a webpage's record can be run against Grapeshot's WordRank™ algorithm (explained elsewhere) to determine the weighted value of the language on the page and determine if there is a match to the keyword segments being used by a partner.

If a webpage is requested but not found in the document store, the URL is sent to the crawler layer, as pictured above, to be crawled, processed and stored in the document store.

Cache Layer ("Slingshot")

While Grapeshot's central document store and categorization layer can handle a large number of requests, speed is of the essence in massive programmatic advertising installations, especially in real-time bidding (RTB) environments.

To enable fast responses, Grapeshot can install a cache of crawled and categorized pages (the product name is "Slingshot") as close as possible to the partner (a.k.a. the customer or client), ideally in the same data center as the partner's servers. A request to this cache returns the categorization and key terms of that page within three milliseconds (3 ms) in most cases.



Each of Grapeshot's partners with a Slingshot installation will have their own dedicated cache of URLs that is built and updated over time. These dedicated caches can contain millions of URLs plus information about the categorization of the textual content of each webpage, which is then matched against the partner's keyword segments.

As pictured in the diagrams above, if a given categorization result is not in the localized cache, a request for that categorization result is then sent to the categorization system, and if not found there, to the crawler to be placed in the queue for crawling. In the meantime, to respond to the request within the required timeframe, Slingshot can return a notice to the partner that information on the requested webpage is not immediately available. Once the page is found in the document store, or crawled and put there -- all of which can happen within seconds or minutes -- the information will be put in the partner's localized cache so the response can be sent within 3 ms the next time it is requested. Pages that have been re-crawled in the categorization layer can be updated in Slingshot installations at a maximum rate dictated by the capacity of the systems supplying the data at the Slingshot location.

---

* Transmission of proper URLs for categorization is the responsibility of our partners (platforms, publishers, etc.) and is outside of Grapeshot's control. Grapeshot does attempt to properly onboard partners, informing and working with them in best practices and proper methodologies. More on these methodologies is below in the "Integrating With Grapeshot" section.

Keyword Segments, Matches and Signals

Keyword segments are another elemental factor at the heart of Grapeshot's technological processes.

Keyword segments (sometimes referred to simply as "segments") are collections of keywords and phrases that, when matched against our categorization of the text of a webpage, will indicate whether that page's epicenter is contextually relevant to that keyword segment. For example, a keyword segment concerning "sports" would contain multiple words and phrases that would indicate – if they appear with enough weight on a page – that the page is about sports.

Each language we cover has hundreds of standard segments covering an array of topics. The segments are constructed by our teams of editors, trained in linguistics and in Grapeshot's processes. Partners can also create, or ask us to create, custom segments to cover topics or niches not sufficiently covered for their purposes by our standard segments. Standard segments can be edited only by our editors, with a formal review process, as described lower in this DoM.

Signals, Matches and Scores

As described in the overview, once a page has been crawled, indexed and categorized, that information is then compared to the relevant keyword segments to determine if there is a contextual match. Matches are scored, to indicate how strong the match is.

Once a match is found and scored, a response is then sent via a "signal" to our partners so they can determine how the page should be treated for advertising purposes. The signal can include other information, as noted below.

Partners set a threshold for the level of match appropriate to their purposes. They can adjust that threshold over time as they wish, for example to surface more pages for bidding to increase their reach, or to increase the level of brand safety, thereby excluding more pages.

Signal Types

Grapeshot signals are provided in standardized alphanumeric formats: a set prefix, descriptive word(s), and the "_" or "-" symbols. The nomenclature is constructed to be easily understandable and differentiated. As one example, a page found to match our "sports" segment will send a response of "gs_sports," as well as a Grapeshot "score" to indicate how strong the match is for that page to the segment. The "gs" prefix indicates "Grapeshot Standard," as described below. There may be sub-segments as well, such as "sports-football" or "sports-tennis" which can be used separately for matching or rolled up into an umbrella segment.

Grapeshot has seven overarching signal response types:

- **gs_ : "Grapeshot Standard"** is for segments designed to be positively targeted, for partners wishing to find advertising inventory on contextually relevant webpages. There are more than 150 signals of this type, and they are translated into all our supported languages.
- **gv_ : "Grapeshot Verified**" is used to indicate a webpage to be negatively targeted (that is, avoided). These responses are for textual content on webpages that contain known brand safety violations or language that is considered unsafe for most brands.
- **gl_: "Grapeshot Language Segment."** This signal indicates the language of a page, for example, "gl_English." This is used to confirm the language is the one desired.
- **gx_: "Information Codes."** Grapeshot gives a "gx_" response when it has not been able to crawl a page and therefore cannot deliver a more meaningful signal. Examples include pages: that have not been crawled and categorized; with no editorial text; that block crawlers; behind logins; or that are temporarily unavailable. These responses can be logged by partners for the purpose of further analysis. A list of gx_ response types is available separately.
- **gq_ : "Grapeshot Page Quality."** This signal is given based on page- or session-measurement data collected through Moat analytics solutions implemented by Moat partners. The signal is used to inform partners of pages that may yield higher levels of users' attention.[1]
- **gs_predicts_: "Grapeshot Predicts."** This is a premium Grapeshot product to help partners find pages that are about trending topics. Whereas standard segments are comprised of a fixed set of keywords that change only if edited, Predicts segments contain a core set of "seed" keywords plus additional keywords that are dynamically updated based on social media and listening technologies.
- **Custom segments.** This is the only segment type that is not part of Grapeshot's fixed taxonomy of uneditable segments, and which is either built by the platform customer or by operators of their sub-accounts (also known as "zones"). By default, these are returned with the convention of *zonename_segmentname*.

[1] Grapeshot is deploying a signal to allow for notification of the findings of Moat-provided testing for IVT and for levels of viewability of an impression.

Determining Match Scores

Grapeshot attempts to score segment matches for categorized pages at the optimum level — one that neither finds too many matches to show true contextual relevance, nor so few that it misses matches that should be found.

To do so, Grapeshot calculates F-scores of standard segment performance against a “gold standard” manually tagged corpus.[1]

For the manually tagged corpus, editors have calculated what they determine to be the correct finding of what each page is about when matched to segments.

The results of Grapeshot’s automated system are compared to the “gold standard” corpus for segment matches, from which a numerical quality score is produced based upon precision, recall and an F-score. (See, too, the DoM document "Maintaining URL Categorization Quality”.[2])

---

[1] https://en.wikipedia.org/wiki/F1_score

[2] Maintaining URL Categorization Quality(see page 13)

Defining Brand Safety Options

Grapeshot offers two standard levels for identifying risks associated with the textual content of webpages:

Maximum Reach: For this level, the textual content of a webpage must be identified as unsafe to be excluded. Pages not identified as potentially damaging are included for targeting. This allows customers to maximize advertising reach while having a standard level of protection.

Maximum Protection: For this level, the textual content of a webpage must be proactively identified as safe to be included. If the page is not identified as safe, it is excluded from targeting (a.k.a., negatively targeted). This protects customers for whom safety is the paramount concern.

Further details are below.

Maximum Reach:

- Pages that are unscanned or unknown are allowed for targeting.
- Pages for which a match is found for standard brand safety (gv_) segments are negatively targeted.*
- Partners can create custom keyword blacklist segments to negatively target further pages for which a match is found.

Maximum Protection:

- Any page that is unscanned or unknown is considered unsafe and negatively targeted.
- A page that has been successfully processed and does not match any of the standard unsafe (gv_) segments* is identified as safe (gv_safe) and offered for targeting.
- Custom safe-from segments can also be added — segments for which if a match is found the page will be negatively targeted.

---

* Customers can choose to positively target segments that would generally be considered unsafe. For example, some advertisers may wish to allow their advertising messages to appear in contexts that are sexual in nature while continuing to negatively target those concerning terrorism.

Mapping to The IAB Contextual Taxonomy

The IAB through its Tech Lab maintains a content taxonomy to help make content classification consistent throughout the digital advertising industry.[1]

Grapeshot has mapped keyword segments to the IAB taxonomy. Our partners can through our interface use the IAB's labeling system, and the appropriate Grapeshot segments are then deployed.

Our aim is to let customers use the IAB taxonomy, developed in consultation with taxonomy experts from academia and industry, in ways that match market conditions and current language usage across every language supported by Grapeshot. Grapeshot monitors updates to the IAB taxonomy and refines relevant keyword segments to conform.

---

[1] https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/

Mapping to The 4A's Advertising Assurance Brand Safety Framework

In September of 2018, the American Association of Advertising Agencies (4A's) Advertiser Protection Bureau (APB) introduced its Brand Safety framework.[1]

The framework lists 13 content categories that, in the words of a 4A's news release, "pose risk to advertisers, whereby advertisers might choose to adopt a 'never appropriate' position for their ad buys."[2]

These 13 categories and the 4A's definition of them are identified in the table below, along with corresponding Grapeshot avoidance categories where available and the ways in which such content is otherwise addressed for webpage textual content.

Mapping of Grapeshot's avoidance categories to the *4A's Advertising Assurance Brand Safety Floor Framework*

4A's Framework			**Grapeshot's Avoidance Categories**	
#	**Category**	**Definition**	**Category**	**Definition**
1	Adult & Explicit Sexual Content	Illegal sale, distribution, and consumption of child pornography Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated	Adult	Avoids mature and sexual webpage textual content
2	Arms & Ammunition	Promotion and advocacy of Sales of illegal arms, rifles, and handguns Instructive content on how to obtain, make, distribute, or use illegal arms Glamorization of illegal arms for the purpose of harm to others Use of illegal arms in unregulated environments	Arms	Avoids webpage textual content around guns and weapons
3	Crime & Harmful acts to individuals and Society and Human Right Violations	Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity – murder, manslaughter & harm to others. Explicit violations/demeaning offenses of Human Rights (eg, trafficking, slavery, etc.)	Crime	Segments within include serious, sex and violent
4	Death or Injury	Promotion or advocacy of Death or Injury Murder or Willful bodily harm to others Graphic depictions of willful harm to others	Death or Injury	Segments within include air, fire, rail, road and sea

Mapping of Grapeshot's avoidance categories to the *4A's Advertising Assurance Brand Safety Floor Framework*				
5	Online piracy	Pirating, Copyright infringement, & Counterfeiting	Download	Relates to online piracy and spam
6	Hate speech & acts of aggression	Unlawful acts of aggression based on race, nationality, ethnicity, religious affiliation, gender, or sexual image or preference Behavior or commentary that incites such hateful acts, including bullying	Hate speech	Avoids derogatory terms including racism, homophobia, and political terms
7	Military conflict	Incendiary content provoking, enticing, or evoking military aggression Live action footage/photos of military actions & genocide or other war crimes	Military	Avoids conflict, war and negative foreign policy webpage textual content
8	Obscenity and Profanity	Excessive use of profane language or gestures and other repulsive actions with the intent to shock, offend, or insult	Obscenity	Avoids webpage textual content that includes offensive terms
9	Illegal Drugs	Promotion or sale of illegal drug use – including abuse of prescription drugs. Federal jurisdiction applies, but allowable where legal local jurisdiction can be effectively managed	Drugs	Avoids webpage textual content related to consumption of drugs, including recreational and performance enhancing use
10	Spam or Harmful Content	Malware/Phishing	Does not map directly to a specific Grapeshot avoidance category, although certain URLs related to this definition may be covered in some part by Grapeshot's "Download" category. Additionally, Grapeshot, as part of our monitoring of keyword segments, manually adds Spam or Harmful Content sites to our internal block-list of pages not to be crawled that will return a "gv_spam_or_harmful_site" categorization when requested by Grapeshot clients	
11	Terrorism	Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society	Terrorism	Avoids webpage textual content around terrorist attacks
12	Tobacco/ eCigarettes/ Vaping	Promotion and advocacy of tobacco and eCigarette (Vaping) & Alcohol use to minors	Tobacco	Avoids all webpage textual smoking content, including vaping and e-cigarettes
13	Sensitive Social Issue/ Violations of Human Rights	Disrespectful and harmful treatment of sensitive social topics (e.g., abortion, extreme political positions, etc.) Acts, language, and gestures deemed illegal, not otherwise outlined in this framework (e.g., harm to self or other and animal cruelty) Targeted harassment of individuals and groups	Does not map directly to a specific Grapeshot avoidance category, although certain text related to this definition may be covered in some part by Grapeshot's "Hate speech" and "Obscenity" categories, and sites deemed inappropriate as noted in line 10, above, may be removed from our crawl list and be designated with a "harmful_site" categorization	

[1] https://www.prnewswire.com/news-releases/4as-advertiser-protection-bureau-delivers-brand-suitability-framework-and-brand-safety-floor-in-move-to-help-advertisers-assess-risk-300717002.html

[2] https://www.aaaa.org/wp-content/uploads/2018/09/APB-Brand-Safety-Floor-Framework.pdf

Use of Metadata

Grapeshot relies most strongly on page text for categorization purposes, and may also include the “title,” "description" and “keywords” metadata fields as part of our consideration process. Generally, we do not rely on page tag data.

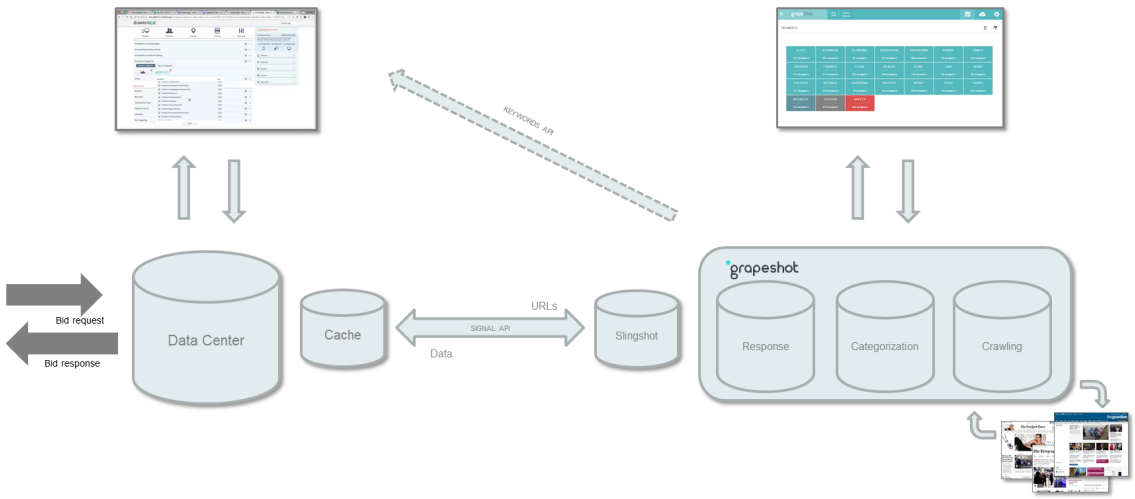
Any systematic tag abuse or failure found to cause miscategorization of pages will automatically be highlighted as an anomaly and be addressed for further editorial review.

Integrating With Grapeshot - The Methods

Integrations Overview

There are three primary ways for partners to to interface with Grapeshot’s servers and exchange data: via API, via flat file and via ad tag.

Via API: We use RESTful APIs. There are two API's: “Signal” and “Keyword.” The Signal API refers to data exchanged via a client’s installation and Grapeshot’s servers to give contextual analysis on page requests. The Keyword API is for handling and updating keyword segments.



As usual for an API, these are sever-to-server (S2S) connections.

Via Flat File: Rather than rely on an API, partners can use a flat file database to specify URLs of pages to scan or block, and determine other parameters on which to act. These flat files can be in standard formats, such as .csv or Excel spreadsheets.

Via Ad Tag: Ad tags*, primarily used by publishers, can be used within an ad placement to interface with Grapeshot’s servers and determine the rules for accepting bids on an ad spot.

---

*Grapeshot's Javascript tag when triggered on a page will execute Grapeshot's page categorization engine and return Grapeshot's page categorization results.

Maintaining URL Categorization Quality

To maintain the quality of URL categorization, Grapeshot's engineering and editorial staff execute the following procedures.

First, an editorial categorization analyst requests at least 1,000 texts from engineering in the specified language.

A software engineer then provides a list of URLs and runs the Grapeshot Signal (gs_ and gv_) against the texts. The top three standard segment matches are noted for each text.

The editorial categorization analyst prepares a new list of texts which is distributed to editors who independently and manually categorize the texts line by line, assigning up to three (gv_ and/or gs_) segments for each text.

The analyst combines the machine and manual categorizations into the corpus for each language and evaluates the precision and recall, the F-Score, and the mean average, error and accuracy.[1]

Errors are addressed as required: Stemmer rules are adjusted, segment terms are updated or, if need-be, a new segment is built.

The above testing process also serves as a check against BM25, an algorithm we use to inform the scoring (a.k.a. weighting) of terms on a page.[2]

Updates to the Crawler for Categorization

In addition to the processes described above, Grapeshot takes action based on feedback from additional sources in order to maintain quality in our URL categorization.

When our monitoring detects technical issues - for example, custom HTML – that hamper categorization of pages on specific domains, we will then take steps to modify our crawler in order to be able to crawl and categorize the epicenter of the pages at that domain.

List of Suspect Domains

Grapeshot maintains a database of thousands of domains on which content or behavior has been detected automatically or by trained staff that are likely to cause brand safety issues.

---

[1]https://en.wikipedia.org/wiki/Precision_and_recall

[2]BM25 (built in part on the work of Grapeshot co-founder Dr. Porter) is a foundational element of information retrieval science and is used by major search engines and others. More about BM25 and its peer-reviewed underpinnings can be found at https://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html.

Maintaining Segment Quality

To ensure the validity of Grapeshot's standard keyword segments, we conduct regular quality assessments of them for each of the languages we cover. Our monitoring focuses first on protecting advertisers from risk from webpage textual content.

There is a multi-step process. First, a team of native speakers based in the market of each language monitors and reviews the terminology being used for each language. Team members have a linguistics background and are trained in Grapeshot's technology and processes. Where there are important variations in a language, such as among American, Australian, British and Canadian versions of English, team members are based in each of these relevant locations so as to monitor and draw distinctions among them.

The frequency of review is determined based upon the frequency of word changes we have discovered in each language. As of August, 2018:

- English is daily.
- French, German, Japanese, Spanish and Chinese (both simplified and traditional) are reviewed monthly.
- Every other language we cover is reviewed every two months.

Our teams to look at the core terms that provide the fundamental evidence of what texts are about and flag evidence of new terms. The teams log and produce reports from a database system that shows the numbers of new terms they discover, what percentage of those terms has been accepted for inclusion, reasons for non-acceptance, comparisons across languages, and graphing along a time series. We regularly check how many new terms we receive and accept so that we can reschedule monitoring frequency if we discover the need to do so.

There are two levels of review. The first is carried out by an editorial manager responsible for reviewing texts, with a second review by a trained editor. When we update the terms for inclusion, we test their validity by using the Grapeshot bookmarklet against regional news articles. We also conduct periodic cross-team reviews to ensure consistency of our methods and processes.

The above is true for Grapeshot's standard keyword segments and for custom segments constructed by Grapeshot for partners. Partners who construct their own custom segments can receive training and instruction in how to do so, however these segments are not subjected to the quality controls of Grapeshot-supervised segments.

Grapeshot is also developing an automated monitoring solution to flag segments for review when quality issues are detected.

Transmission of Changes

Data used in our processes are kept reliable as to current conditions, with variance of below five percent. Here is one example to illustrate our process: The phrase "food porn" was found to be causing pages to be incorrectly flagged as adult content. This was addressed by our in-house editors, who made changes to the relevant segments, pushing them live immediately. From then on, requests for URLs containing that phrase were treated correctly.

Because Slingshot installations have a recommended TTL of 15 minutes before re-querying, they may in a very few instances for a short time use the previous version of a segment that has been edited. Caches that are owned and supervised by clients are outside of our control.

Detecting Verification Avoidance and Content Variations

There are instances in which the epicenter of pages under identical URLs could possibly be found to vary due to a number of circumstances. Those circumstances include:

- **Verification Avoidance**. There may be instances in which “black hat” operators in the ecosystem wish to show one version of a page to Grapeshot’s crawlers but another to users who come to the site. This kind of spoofing could be done in order to allow ads to be served onto a page that would otherwise be blocked after having been identified as being contextually unsafe.
- **Variance in Geographic Location**. Some sites are blocked from being accessed from certain geographic locations. (For example, sites identified as facilitating the pirating of content cannot typically be accessed via certain European ISPs.) In other instances, pages could vary according to a user's location, so that users in different locales see different versions of a page under the same URL.
- **Desktop vs. Mobile Variations**. There may be variations in pages served to users visiting a site on desktop vs. mobile devices.
- **Javascript Execution**. Javascript code is sometimes used to serve content on a screen, generally for mobile devices. There are also situations in which text and other content is served to a mobile screen after a user executes a command, such as by clicking a "read more" button to see a full page.

In order to test such variations, Grapeshot configures virtual machines containing the crawler technology on a smaller scale than the large scale production crawler infrastructure. These virtual machines can be deployed anywhere in the world and can vary the crawler dimensions – user agent string, geography, device type, Javascript – to test their effect on categorization. *This process is used for testing only*, as Grapeshot wishes in its at-scale production operations to employ best practices, such as transparently identifying its servers as coming from Grapeshot.

Should intervention be deemed necessary due to meaningful levels of variance, Grapeshot will intervene manually and craft counter-strategies. More details are below.

Monitoring of Pages Grapeshot Cannot Crawl

There are times when Grapeshot is unable to deliver a meaningful signal about a page, such as when the page has no editorial content, is behind a login, is temporarily unavailable, or when Grapeshot's crawlers are blocked from a domain or sub-domain.

In such an instance Grapeshot will return an "Information Code" signal to our partner (via a "gx_" preface in the signal).

It is then up to the partner's discretion whether to include the page for advertising bids, such as if the page is at a known and trusted domain or is, based on previously gathered evidence, believed to be safe. Partners, of course, also have the option to block such pages from consideration for advertising bids.

Grapeshot is also deploying processes for evaluating pages for which there is an abnormally high proportion of gx_ responses (compared to observation of the marketplace in general) or that have unexplained and significant shifts in the proportion of gx_ responses returned.

In mobile apps we have limited capabilities for contextual analysis due to limitations on the ability to look into pages within many apps. We are able to glean information from the publicly available app descriptions and to generally gauge potential brand safety issues from PEGI scores, where available.[1]

---

[1] https://en.wikipedia.org/wiki/Pan_European_Game_Information

Measures to Detect Verification Avoidance

There is at least a theoretical possibility that a publisher detecting a Grapeshot crawler may try to fool our system by delivering content different from that which a user would see when he or she visits the page. Such a publisher would be considered a nefarious player who is probably trying to allow ads to be served onto a page whose proper contextual analysis would identify it as a likely candidate for exclusion for brand advertisers. Grapeshot implements quality control procedures in order to detect such scenarios. We have not, to date, found widespread or systematic variation in pages beyond normal editorial updates.

Addressing Possible Inaccuracies Due to Geographic Location

Access to URLs can sometimes vary by geography. Some publishers' pages, for example, are systematically blocked from being viewed in certain countries by those countries' ISPs. Grapeshot crawlers that are identified as coming from those geographies would, thus, also be blocked from crawling.

It is also theoretically possible that users from different geographies could be shown different content under the same page URL.

Grapeshot has deployed empirical testing to determine URLs for which such geographic blocking may be in use and to date has found no such variations.

Variations Between Desktop and Mobile Versions

Grapeshot has implemented empirical tests to determine variations among the desktop and mobile versions of the same pages.

We are, for example, testing the "m." URLs offered to mobile device browsers to see how they differ from the "www." desktop versions.

In other cases, notably in Asia, content is served onto pages via execution of Javascript rather than through HTML, primarily for mobile devices. Grapeshot in such cases executes the Javascript to pull text onto the page for crawling and analysis.

In other instances, content on a mobile screen is loaded onto a page via a user-initiated button, such as by clicking “read more." Grapeshot currently does not execute these buttons, as there can be a plethora of buttons on any given page. We are in the process of determining if there is a methodology for distinguishing among these buttons, so we might execute only those that are relevant to Grapeshot's processes, thereby neither taxing publishers' servers and causing potential latencies, nor incurring large burdens for our clients.

Analysis is Absent Of User Information

Grapeshot crawls at the webpage contextual level (in a similar fashion to search engines) after receiving a request to analyze the context of a page.

User information (such as may be transmitted by cookies or browsing history) is not collected as part of our analysis.

Grapeshot's systems will therefore not capture variations in page epicenters that could possibly occur based upon user profile variation.

A Word About Cookies

Grapeshot does not rely on cookies or information gleaned from them. We do not collect, store or use information about an individual consuming content we analyze. This makes us different from a large proportion of the participants in the advertising and publishing ecosystems.

There are a few instances in which our systems may interact with cookies to perform our work:

To Access a Page. Many publishers block users from viewing their pages until the users click an “opt-in banner” to indicate acceptance of a publication’s privacy policy. A cookie is then placed in the user’s device, so the pages can later be accessed without hindrance. Such opt-in banners can block Grapeshot’s crawlers. In order to gain access, Grapeshot may therefore receive and store a cookie so we can crawl the pages and begin contextual analysis. Such cookies have no relation to any user.

To Match Contextual Information With User Data for Analysis. A small number of Grapeshot partners matches Grapeshot’s contextual analysis of pages with their own cookie data on which users have visited those pages. They can then use that information for ad targeting purposes. Grapeshot does not have or retain any of the cookie data, nor do we use it in any way in our systems.