ORACLE
Data Cloud

Oracle Contextual Intelligence

Description of Methodology

December 2020

Oracle Contextual Intelligence Description of Methodology

Executive Summary

This Description of Methodology (DoM) is a summary of processes employed for the delivery of the Oracle Data Cloud Contextual products and services. It includes a general description of the products' 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

Contextual Intelligence, at its core, is about contextual understanding of content, processed at the massive scale and speeds required by automated advertising technology.

Our technology crawls pages then uses information retrieval processes to determine they core content of the page. We then compare that content to “contextual segments", sets of carefully compiled words and phrases concerning specific topics, to determine if there is a match, and how strong the match is. Our customers and partners use this information to report on the contextual relevance of impressions bought/sold; to make monetization decisions (for media owners); or (in a programmatic buying environment) to indicate whether a page should be negatively targeted (removed from consideration) or be positively targeted.

Our technology works for web pages on desktop and mobile, and we are implementing the technology for video by converting speech to text, then applying our contextual segment matching processes. At present we are exploring the incorporate of image recognition and detection as part of the classification logic, but this has not been formally included in our production systems. Within the assets ODC crawls and categorizes, the system does not currently evaluate the context for in-page advertisements, side-bar content, or external links. As a consequence, ODC contextual intelligence is considered by the MRC to meet the requirements for 'property level' brand safety[1].

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, IAB and TAG, we have deployed multiple processes as described below.

Primary Users and Use Cases

The dominant use of Contextual Intelligence technology and processes is to detect the important meaning of the core content on a page, then offer that understanding for the purpose of enhancing

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

both brand safety and ad targeting. Our aim is to offer media buyers, ad tech partners, and media sellers (a.k.a. publishers) a transparent lens around content.

Contextual Intelligence is used by both the buyers and sellers of digital advertising. Brand marketers and the agencies representing them use the technology via the programmatic buying platforms they employ, as do publishers wishing to better serve those advertisers, increase revenue and serve their audiences. Technology platforms may include DSPs, SSPs, advertising exchanges, and measurement and verification services.

Contextual Intelligence technology is employed on the web — desktop and mobile — and for mobile apps, for text and video content. For every platform and use-case, the fundamental technology is the same.

Note that all of these constituents use ODC for categorization and ranking of content on a page. Other platform functions, such as ad-serving, viewability verification, identification of invalid traffic (IVT/SIVT), measurement of audiences and other cookie implementations are not performed by the Contextual Intelligence system. Predictive viewability and IVT targeting derived from those measurements is delivered via the same integrations as Contextual Intelligence but are outside the scope of this document.

Use Cases

There are two primary business uses for Contextual Intelligence technology:

- Brand Safety. ODC gives advertisers the opportunity to avoid serving their messages adjacent to content they find objectionable.
- Targeting. Advertisers and publishers use ODC to find pages with content that is contextually relevant to advertisers.

The overarching goal of Contextual Intelligence technology is to get at the important meaning of the core content and understand it in the desired context for the specific application by applying the appropriate ranking and levels of importance to the various terms. 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 in certain advertising circumstances.

ODC also allows custom gradations and control. Some brands, for example, may not object to having their advertising message on an article or video in which the word “ball” is used in even its racy sense, or may be fine with ads placed on the page if the named activity is not the focus of the story.

Contextual Intelligence Processes

Overview

To do its work, ODC Contextual Intelligence:

1. Crawls hundreds of millions of pages daily and indexes their core content (textual or other). Algorithms are used to identify the relative "weight" of all words within content (e.g., a news story on a web page or transcription of a news program). These weighted words are generated as a fundamental "atomic" composition of that page asset.
2. Separately creates groups of words and phrases known as "contextual segments," which are themselves sets of unambiguous words determined to be reflective a particular topic. The set of words in a 'contextual segment' are described as the 'contextual segment definition'
3. Compares the weighted atomic composition of the processed content to the contextual segment definitions and provides scores to indicate the degree of similitude, known as a 'probabilistic match'. Contextual Intelligence will make multiple probabilistic matches between a set of such contextual segments and the document.

There are other variations and enhancements of this core categorization process that utilize different kinds of contextual segments. These might utilize some of the infrastructure of Contextual Intelligence, or incorporate additional mechanisms in addition. Two variant targeting options are summarized here:

- Predicts, which generates segment definitions based on trending content, provides a dynamic/changing segment definition useful for targeting. *(Uses core contextual intelligence, plus a mechanism to adjust segment definition according to recently categorized content.)*
- Page viewability and invalid traffic data from ODC's Moat Analytics engine is also used to create 'predictive' segment definitions which describe URLs/assets in terms of likelihoods to be viewable or invalid/fraudulent, aka 'Prebid Viewability' or 'Prebid IVT'. *(Only uses Contextual Intelligence delivery methods; does not utilize contextual categorization components.)*

The final output of the Contextual Intelligence process is effectively a set of contextual categories associated to a given asset (e.g a webpage). Advertisers (and those servicing them) use the matched context segments associated to each webpage/mobile app/video in order to make decisions about avoidance (brand safety) or targeting.

The Science Behind ODC's Contextual Intelligence

ODC'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 (ODC's contextual tech acquisition), focused 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 Google, IBM and Microsoft.

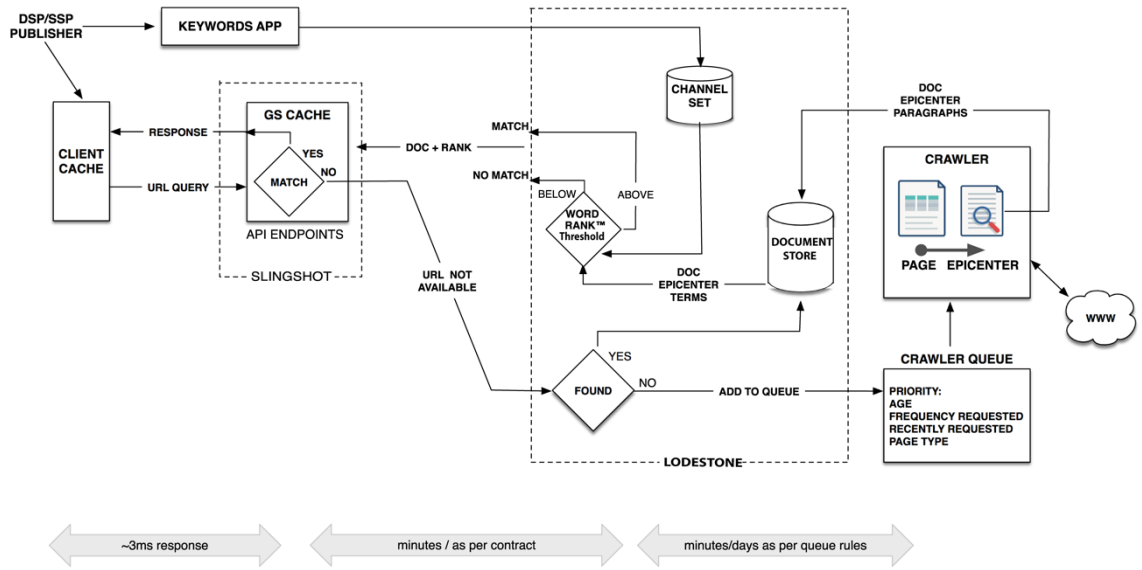
Dr. Porter developed new ways of not only processing and understanding language on a page but also doing so quickly and with minimal intrusion into the workings of the page. Dr. Porter's vision for Grapeshot was to determine the full meaning of a document and then to apply it effectively on a massive highly dynamic body of content like the Internet. Contextual Intelligence at ODC 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 of the semantic solutions which must utilize advanced machine-learning processing to model and generate a set of rules to a core system, ODC's two-step process – processing the page then matching against

contextual segments -- may be considered more flexible and adaptable, with greater degrees of human input.

Contextual Intelligence Architecture and Flow

The Layers: Crawler, Categorization and Cache

ODC's Contextual Intelligence crawls and contextually analyzes content 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. ODC 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 Contextual Intelligence infrastructure.

1. Crawler layer (right): a system that manages crawling of the core (the main HTML, also known as the "epicenter") of requested pages. 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 as described earlier
3. Cache layer (known as "Slingshot"): a localized installation set up to enable distributed segment matching and support scaling and low-latency requirements.

Crawler Layer

The request for a page at a given URL can originate from a number of sources but typically will be have been from a partner's technology platform or server that handles ad serving, measurement and/or verification.[2] Pages are crawled at a rate of over 500,000 per minute and are re-crawled to check for changes depending on the propensity of a particular page at the given URL to change content.

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 website is used for all partner implementations. To account for pages that are the same content but have subtly different URLs -- such as from parameters toward the end of a URL that are caused by web analytics software -- Contextual Intelligence strips out those parameters.

Sites that ODC is attempting to crawl can exclude or block the crawlers by various means, such as via their robots.txt page or by specifically excluding ODC's crawler. If ODC is unable to crawl a page, information about that inability will be indicated to ODC's partner or customer. (There is further information below on crawler limitations and scenarios ODC has explored for hypothetical scenarios of an entity attempting to thwart or deceive crawling.)

Epicenter, Adjacent and Non-Text Content

Once the Contextual Intelligence crawler has received a request to scan a page, it finds, crawls, and downloads the "epicenter" of that page (that is, the core content) from its HTML -- not the CSS, JavaScript, images, navigation, footer, and other areas tangential to the main meaning on the page. (Picture a typical news page. ODC's technology will download the central elements of that page but not the side elements which may include "Related Stories," additional linked headlines, and so on). As mentioned before, 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, the MRC currently denotes Contextual Intelligence as a 'property-level' solution.

That said, we are aware that surrounding and non-text content, including images and content that may be provided for by JavaScript executions, can be seen to affect the context of a page as presented. ODC has been investigating methods for incorporation of image detection into both video and textual classification, though they are not production ready at this time. As industry best practices are developed and adopted, we can accordingly institute technologies to scan additional components on a page. This may require implementation of a much more intensive level of crawling

---

[2] * Transmission of proper URLs for categorization is the responsibility of our partners (platforms, publishers, etc.) and is outside of ODC's control. ODC does attempt to properly onboard partners, informing and working with them in best practices and proper methodologies.

and analysis, causing greater server loads, and perhaps affecting page latency as well as customer pricing.

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[3] request. (ODC 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 Contextual Intelligence contextual segments.)
2. Normal requests as described above
3. Refreshes of stale pages whose time-to-live (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 30 minutes. 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.

ODC conducts empirical tests to assure that the 30-minute minimum threshold is sufficient to detect content epicenter changes at a rate of 95% or above. Should the tests find meaningful content 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 page is crawled, its information is sent to the Contextual Intelligence categorization layer (which we call "Lodestone"), where the page's information is kept in a corpus -- a central data store of the information from all crawled pages. The Contextual Intelligence system's central data store holds more than 8 billion documents at a time, and is an ever-growing, frequently updated record of all pages that have been crawled.

From the corpus, the page's record can be run against ODC's contextual intelligence algorithm (explained elsewhere) to determine the weighted value of the words on the page and determine if there is a match to the contextual segments being used by a partner.

If a page is requested but not found in the corpus, the URL is sent to the crawler layer, as pictured in the top figure, above, to be crawled, processed and stored in the corpus.

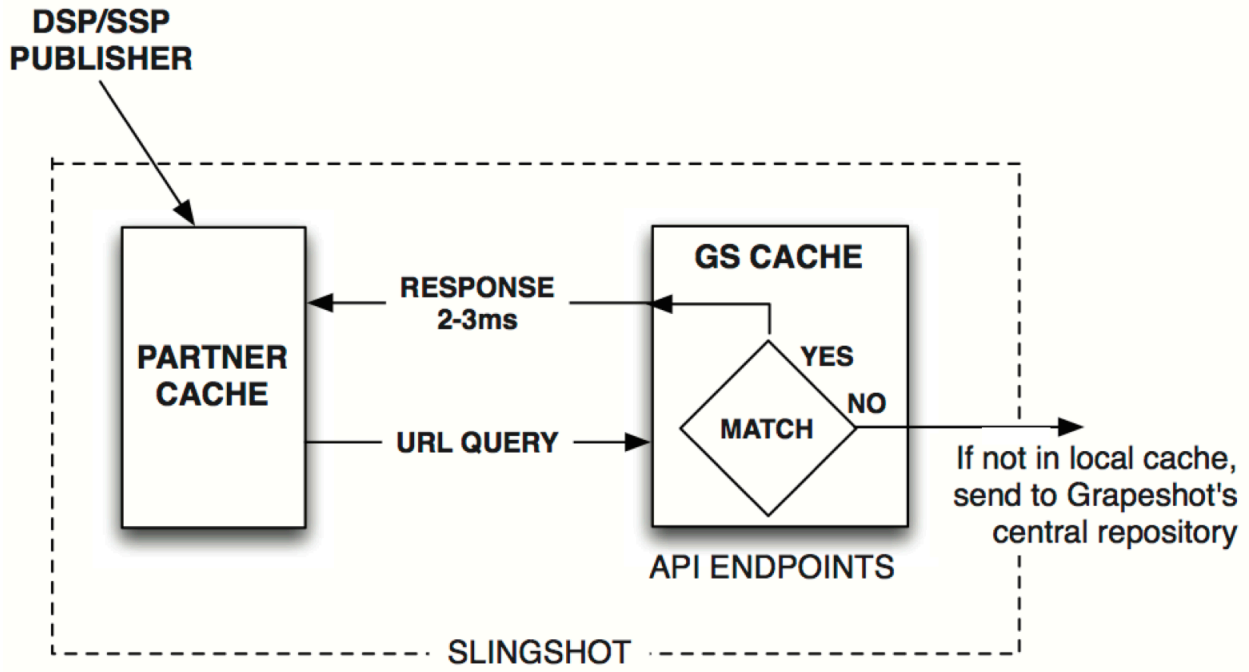
---

[3] ODC utilizes a 'quick check' browser-based tool that lets editors, developers, and customers view single-page classification data.

Cache Layer ("Slingshot")

While the Contextual Intelligence central corpus 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, ODC often install a distributed 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), commonly in the same data center as the partner’s servers. A request to this cache returns the key terms on the page and the categorization of that page, within three milliseconds (3 ms) in most cases.



In their installation, each of ODC's partners 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 each page, which is then matched against the partner’s contextual 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 page is not immediately available. Once the page is found in the corpus, 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.

Content Crawling Risks and Limitations

Introduction

The Contextual Intelligence crawling system has been developed over a number of years and whilst it is intended to have good coverage and handling for the vast majority of sites there will inevitably be some sites and pages that cannot be correctly or fully processed. New sites appear all the time that may or may not render in compliant and normal ways.

Geographic and site access restrictions may also limit the crawler's ability to extract text from sites.

The intention of this document is to highlight any known issues or likely limitations.

Content Extraction Limitations

These are known issues and risks. As part of the support process when incorrect text extraction is reported tickets are raised against the ODC development team to try and address them. In some cases, generic fixes can be applied, others require bespoke fixes for the publisher/site/page/asset in question.

The text extraction approach is fairly tolerant of html mark-up and does not rely on publishers indicating article content, instead looking for dense textual content, an approach that is, in general, agnostic to actual language. It is rarely, if ever, necessary to do language specific support.

Complicated layout schemes

Unfortunately, some sites use an extremely complicated and large html mark-up, including heavy use of CSS to layout the page properly. This is obviously not good practice in general, and also makes accessibility for visually impaired users very poor, so these are not usually, almost by definition, mainstream sites. Without fully rendering the site in a supported browser it can be hard to determine where the actual main article text lies, especially if the mark-up is so broken that it only works on browsers as they have significant on-going developer effort to be very tolerant of such broken html.

Dynamic rendering

A small number of publishers choose to render their sites dynamically. Typically, this takes one or two forms. Either the payload is part of the site html payload but hidden within some embedded content and rendered dynamically with JavaScript as the page loads. Or the initial payload kicks off an extra call, or calls, to the site to obtain the actual content. Without fully rendering the page, which is very expensive at scale, it’s not practical to get the correct content view, especially for the latter case.

However, in some cases it is possible for the text extraction algorithm to extract the embedded content or add specific rules to do so. It’s also possible in some cases to add rules to go straight to the actual article download request and bypass the loading wrapper. This latter tends to be a on a site-by-site basis as it is a bespoke handling in general.

Assessments ODC has done on these two cases suggest that the number of such URLs is a fraction of 1% percent against the whole of the internet. Rendering like this is generally going to cause issues for search engines too, which is not good for sites that need visitors for their business model.

There are some known sites that use these sorts of rendering models specifically for their home pages, actual article pages are presented in a more normal fashion. Home pages have their own associated risks and limitations, as discussed later.

Infinite Scroll design

Beginning in the middle of the last decade (2010s), a number of website publishers chasing higher article views opted for a continuous scrolling format to their article pages. In these the last screenful of content from an article would cause the load of the next article from the section or site. So, as the user scrolled, they were presented with endless content.

As our crawler system does not render the page or run any JavaScript the Infinite Scrolling nature of the pages presents no challenges for the crawling system. Each article is processed individually as content in the usual method.

In almost all of the implementations Context has encountered, the URL was changed as the new article was loaded. This new URL is used in the requests for adverts to be loaded for the next article. Where this happens, context is able to crawl and classify each article on its own URL and the infinite scrolling causes no issues to our crawlers or classification service.

Throughout this phase in website design ODC's Contextual Intelligence team is aware of only a small handful of publishers globally (<10) that didn't update the given URL when requesting adverts for the newly loaded content. In these cases, the classification will continue to be based on the original URL as we wouldn't be aware that the content of the viewed page had actually changed.

Mixed Language

This is more an issue with APAC sites, where from time to time a mixture of English (or sometimes other, but mostly English) and the primary APAC language appears. The Contextual Intelligence system can only operate using a single language, although the segment definitions do support using English words in non-English segments. When determining the language of a site the language that is most dominant is typically used. On occasion English may be the dominant language, especially if the text is quite short, which could lead to incorrect text being extracted as this detected language is used as part of the scoring mechanism for determining the textually dense part. This may also impact categorization where a less ideal language and its segments are used.

Ambiguous Body/Article

Some pages can have multiple blocks of text but possibly split into multiple distinct parts in the html mark-up. In general the text extraction system will pull all the text, but if the blocks are sufficiently distinct it is possibly that only one of them will be selected, and selection is based on an algorithmic scoring model looking for the most interesting text. In some rare cases the chosen block may not be the best one to represent the page, some sites put a large footer block that looks like an article to the

algorithm for example. Whilst things marked up as footers in some general way are suppressed by the algorithm not all publishers give any indications that footers are of this type.

Home Pages

This can be the most complicated version of the previous type. Publishers often put large numbers of snippets of articles on home pages which are all logically distinct text blocks. In general, where possible, Contextual Intelligence extracts all the text on home pages, rather than trying to find one dominant text block. However, since this then results in the total extracted text covering multiple topics categorization may not be as meaningful as for a single article. The scale of this gap is minor in terms of the total proportion of unique page requests sent to the Contextual Intelligence system, but home pages can represent a much larger proportion of individual page impressions served. However few of ODC's Contextual Integrations are at the impression-level (described later in a section regarding caches and latency) and as such partners are not utilizing the most current Home Page classification possible, regardless of how ODC categorizes.

Site Access Limitations

For reasons beyond ODC's control it may not be possible to access the text context on some sites and pages. A number of the known cases are listed here.

Paywall

Some sites implement a paywall system as part of their business model. However, these fall into multiple categories.

Some publishers have allowed the Contextual Intelligence crawling systems to bypass their paywall (usually after discussion with ODC or possibly proactively) as they recognize the value added by this.

Some sites allow a limited amount of reading before restricting access. In general these can be processed by Contextual Intelligence as normal because of the implementation ODC uses for accessing them. As this is the easiest model for publishers to implement this is quite a common one but poses no risk to Contextual Intelligence at all.

Other sites will show a leading snippet of the article with a subscription link to see more. Since there is no distinct and machine-readable way to know this has happened in general there is scope for only this snippet text to be considered for categorization. This could potentially result in a mis-categorization. However, note that even here some of these sites do in fact have the entire content in the page, it’s just being hidden, so Contextual Intelligence may in fact have the entire text anyway.

Login required sites

Arguably a more extreme version of the Paywall, all content requires a valid login to see it. This usually results in a request to view the content causing a redirect to the login or register page. Contextual Intelligence will recognize this in general and not follow the redirect or try extracting text. However, some publishers may make this quite opaque and hard for an algorithm to

determine the actual content is not being shown, but links to a login or registration are being provided instead, this is quite rare, best and most common practice is to redirect.

ODC never seeks to get login details for such sites; if the publisher declines to allow access without a login the text content will not be available for this site.

Content with restricted access

A continuation of the login required site; some sites may have private content only viewable by specific logged in users, including for example, email or social media. The Contextual Intelligence system would be unable to view this content under any circumstances and would be unable to provide categorization for it. ODC would never seek to gain access to this type of content via specific login but would look to create direct API-type integrations.

IP based/Geographic restrictions

Usually for legal reasons some sites restrict which parts of the world can see their content. Some US sites, for example, decline to serve any content to EU nations as they believe GDPR regulations make this untenable. Other sites may be providing a service that for other legal or commercial reasons cannot be made available outside its primary geography. Streaming sites are one example.

Context Intelligence page crawling systems are hosted in the UK, and as such the IP addresses are marked as UK addresses in geographic databases.

Such sites would typically redirect to a holding page stating the restrictions, for which Contextual Intelligence would not extract content. However, in some rare cases publishers choose to simply replace the actual content with the restriction notice, and there is a risk that ODC's systems will use this text for categorization, as it has no way of knowing this is not the real text. Fortunately, such cases are rare, publishers rapidly find that search engines also run into this and it dramatically reduces their visitor numbers.

Against the broad landscape of the internet, such sites are very rare. However, ODC plans to implement, at least, US based crawling systems to minimize the impact of this in the near future (target calendar year 2021).

Geographic Content

Most common for a news publisher, content may be tailored to the determined geographic location of the request. As Contextual Intelligence runs from the UK at this time it's possible that UK content is returned as compared to US user for the same page, leading to a different categorization. However, the vast majority of publishers do not actively change content in this way, as instead they run multiple sites on different domains for each geography. Options are presented to the user to select which geography they want. Full articles are, in general, able to be processed in the correct way by Contextual Intelligence.

The exception, in some cases, is the home page. On some sites visits to the home page always cause a redirect to the geographic home page, visitors from the "wrong" geography are simply denied

access to that geography's home page. This is a problem for Contextual Intelligence as it may never be possible to extract the home page content from all geographies. Fortunately, such sites are very rare, as most publishers offer this is an option, rather than forcing.

In some cases, ODC has been able to either request the publisher allow multiple geography home pages to be seen or has been given a way to override the forced move.

Robots.txt limitations

A standard machine reading protocol, robots.txt, can be used by sites to allow or deny access to content. A well-behaved page crawling system, such as ODC's Contextual Intelligence, must obey rules defined here. Publishers can define request rate limits or allow/deny entire site or parts of the site from crawling systems. If ODC is asked for categorization for such a denied page, it will not be possible to extract its text.

It is also possible that the publisher has set the rate limit so low that getting good coverage of the site's content is very difficult. If a site has 1000's of pages but the publisher only allows 10 pages to be downloaded per day, which some do, it's obviously not going to be possible to keep up.

Rate limit restrictions

Publishers, and their systems, are often wary of being exposed to large scale requests, denial of service attacks in particular. Many have, or use, systems that try to detect potential abuse, too many requests in a short space of time can result in subsequent requests being blocked for a period of time. ODC tries hard to avoid this, utilizing automated rate limiting strategies within our crawler system.

Note that as a well-behaved crawling system, Contextual Intelligence does advertise in the request to the site that it is a crawling system, no attempt is made to hide this or pretend to be a human. Some sites do actively block all crawling systems, regardless of purpose.

Requests to desist all crawling

On rare occasions publishers decline to use the robots.txt system and choose to ask ODC to desist from crawling content instead. ODC always abides by these requests, and a blocklist is maintained for this purpose. This is a relatively small list as the majority of sites that run adverts recognize the need for content analysis.

Text is not available for such sites.

Faked Content

Since ODC advertises in its crawl requests that it is a crawling system, there is theoretically scope for a publisher to return different content as a way of controlling the resultant categorization. However, as described below, ODC has run controlled studies on this and has yet to find any indication that this happens in practice.

Detecting Verification Avoidance and Content Variations

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

- Verification Avoidance. There maybe instances in which “black hat” operators in the ecosystem wish to show one version of content to ODC'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 containing unsafe content.
- 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, content 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 content 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 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, ODC 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 of content. *This process is used for testing only*, as ODC wishes in its at-scale production operations to employ best practices, such as transparently identifying its servers as coming from Oracle Data Cloud.

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

Other Crawling / Categorization Notes Limitations

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

- User-based Categorization Variance
- Desktop vs. Mobile Variations
- Verification Avoidance

In order to test such variations, ODC 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 of content. *This process is used for testing only*, as ODC wishes in its at-scale production operations to employ best practices, such as transparently identifying its servers as coming from Oracle Data Cloud.

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

User-based Categorization Variance

ODC crawls content at the page level (in a similar fashion to search engines) after receiving a request to analyze the context of content on page. User Information (such as may be transmitted by cookies or browsing history) is not collected as part of our content analysis and the Contextual Intelligence systems will therefore not capture variations in content that could possibly occur based upon user profile variation.

However, 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 ODC's crawlers. In order to gain access, ODC may therefore receive and store a generic non-user-based cookie so we can crawl the pages and begin contextual analysis. Such cookies have no relation to any user.

A Word About Cookies and Downstream Usage of Context

Oracle Contextual Intelligence systems do not have or retain any of the cookie data, nor do we use it in any way in our system processes. 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. A small number of ODC partners (e.g. publishers) may match Contextual Intelligence analysis of pages with their own consented cookie data on which users have visited those pages. They can then use that information for ad targeting purposes.

Mobile vs. Desktop Variance

Oracle Data Cloud has implemented empirical tests to determine variations among the desktop and mobile versions of the same pages. We are, for example, testing the "m." prefix; 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. ODC in such cases executes the Javascript to pull the content onto the page for crawling and analysis, as constrained by dynamic rendering limitations listed elsewhere.

In other instances, content on a mobile screen is loaded onto a page via a user button, such as by clicking “read more." ODC Contextual Intelligence 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 ODC's processes, thereby neither taxing publishers' pages and causing potential latencies, nor incurring large burdens for our clients.

Measures to Detect Verification Avoidance

In crawling pages, ODC will in the large majority of cases identify its crawler as coming from ODC Contextual Intelligence (Grapeshot). The technological means of self-identifying ODC's crawlers is via the user agent string, a line of text that identifies the crawler to a web server it is visiting as part of the protocol that requests the page. There are multiple reasons for doing so, primarily that ODC

desires to be a good netizen, and transparently identifying oneself on the Internet is considered a best practice.

However, there is at least a theoretical possibility that a publisher detecting an ODC Contextual Intelligence 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 website would be considered a nefarious player who is probably trying to allow ads to be served onto a page whose true content would identify it as a likely candidate for exclusion for brand advertisers. To detect such scenarios, ODC crawls a sample of pages1 from websites without identifying the crawler as coming from ODC. We then compare those sampled pages to the same pages offered to ODC-identified crawlers.

We have not, to date, found widespread or systematic variation in pages beyond normal editorial updates but continue to run tests. Should we find such variations on any publishers' sites, we shall implement a methodology to identify those publishers to our clients, and offer them the capability to exclude those publishers' URLs from their ad serving.

Contextual Segments, Matches and Signals

Contextual segments are another element at the heart of ODC's Contextual Intelligence processes. Contextual segments (sometimes referred to simply as "segments") are collections of words and phrases that, when matched against our categorization of a page, will indicate whether that page's content is contextually relevant to that contextual segment. For example, a contextual 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 used for both targeting and brand safety purposes. The segments are constructed by our teams of editors, trained in linguistics and in ODC's Contextual Intelligence 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 contextual segments. Standard contextual segments can be edited only by our editors, with a formal review process, as described later in this document.

There are other segment types such as viewability and invalid-traffic predictors which are not matched using the categorizer's algorithms but, where available, are included in the response.

Matches

As described in the Process overview above, once a page has been crawled, indexed and categorized, that information can then compared to the relevant contextual segments to determine if there is a contextual match. Each time a request for categorization is made by a partner/client, the response reflects one or more matching segments that have been identified in run-time. In some cases it is possible for clients to also receive the words/phrases that prompted each match, though this is only on a case-by-case basis. The response can also include other information, as noted below.

Scores

At the moment of request, all potential contextual segment matches are scored within Contextual Intelligence to indicate the relative strength of each potential match *at that moment*. While there may be many contextual segment matches, ODC sets a threshold for the level of match to maintain consistent results. Segments which fall below this threshold match score are not part of the response to the categorization request as the 'contextual signal' are considered to be too 'weak'. Scoring is used only for internal purposes; it is scoped largely on each request for that client and is not made available to customers.

Signals and Signal Types

The collective response for a given page/asset request can be referred to as the 'contextual signal'; it is a composite of the various categorizations generated by the Contextual Intelligence system. All 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 that contains content that matches our "sports" segment will send a response of "gs_sports". (The "gs" prefix, still used in our internal systems, stands for "Grapeshot Standard," as described below.) There may be sub-segments as well, such as "gs_sports_football" or "gs_sports_tennis" which can be used separately for matching or rolled up into an umbrella segment.

ODC's Contextual Intelligence has six overarching signal response types (using legacy 'GS' designations to reflect the original company Grapeshot'):

- **gs_ : "Grapeshot Standard"** is for segments designed to be positively targeted, for partners wishing to find advertising inventory on pages that contain relevant content. There are more than 150 signals of this type per language, and they are translated into all our supported languages. These contextual categories are the same for any integration.
- **gv_ : "Grapeshot Verified"** is for brand safety responses, to indicate a page that is to be negatively targeted (that is, avoided). These responses are for pages that contain known brand safety violations or unsafe content 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 one desired.
- **gx_: "Information Codes."** ODC's crawler system will generate a "gx_" response when it has not been able to crawl the content on the page and therefore cannot deliver a more meaningful signal. Examples include pages: that have not been crawled and categorized; with no editorial content; 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 Quality."** This signal is a predictor for likelihood to be viewable or invalid given based on that page- or session-measurement data collected through the Moat analytics platform. The signal is used to inform partners of content that may yield higher levels of users' attention.[1]
- **gs_predicts_: "Grapeshot Predicts".** This is a premium Contextual Intelligence product to help partners find pages that contain content about trending topics. Whereas standard contextual segment definitions are comprised of a fixed set of words and phrases that change only by ODC editorial teams, Predicts contextual segment definitions contain a core set of words and phrases that are dynamically updated based on the feed of the latest pages classified by the ODC Contextual Intelligence system, reflecting the trending consumed content.

- **Custom segments.** This is the segment type that is not part of ODC's fixed taxonomy of uneditable segments, and which is either built by the platform customer or by operators of their sub-accounts (identified as "zones"). In order to differentiate these from standard segments, these are returned without any "gs_" prefix using the convention of *zonename_segmentname* instead.

Determining Match Scores

The Oracle Contextual Intelligence categorization engine 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, ODC calculates F-scores [4]of standard segment performance against a "gold standard" manually tagged corpus. 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 the categorization 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.

Defining Brand Safety Options

Oracle Data Cloud offers two standard levels of brand safety:

Maximum Reach: For this level, a page must be identified as unsafe to be excluded. Content that is not identified as potentially damaging is included for targeting. This allows customers to maximize advertising reach while having a standard level of brand safety protection.

Maximum Protection: For this level, a page 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:

- Content that is unscanned or unknown is allowed for targeting.

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

- Content for which a match is found for standard brand safety (gv_) segments is negatively targeted.[5]
- Partners can create custom keyword blacklist segments to negatively target further pages for which a match is found.

Maximum Protection:

- Any content that is unscanned or unknown is considered unsafe and negatively targeted.
- Content that has been successfully processed, matches at least one standard (gs_) segment and does not match any of the standard unsafe (gv_) segments[6] 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 content will be negatively targeted.

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.[7] Oracle Contextual Intelligence has a full set of segments exactly matching the content categories published as the IAB Content Taxonomy (v2.2), as of January 2021. 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 Oracle Data Cloud. We wish to give partners the ability to use the IAB taxonomy with a breadth of reach, targeting and brand safety to match current conditions and language usage. Oracle Data Cloud monitors updates to the IAB taxonomy and refines its taxonomy to conform.

[5] 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 adjacent to content that is sexual in nature while continuing to negatively target content concerning terrorism.

[6] See #4 above

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

Mapping to the 4A's Advertising Assurance Brand Suitability Framework and Brand Safety Floor

In September of 2018, the American Association of Advertising Agencies (4A's) Advertiser Protection Bureau (APB) introduced its Brand Suitability Framework. In early 2020, the Global Alliance for Responsible Media (GARM) under the World Federation of Advertisers (WFA) joined this framework and agreed to a set of eleven categories as a new standard of safety. The IAB was asked to steward the development of this framework into a workable set of standards around which 3rd party vendors could align safety categorization.

The eleven categories create a 'floor' whereby advertisers might be able to choose to adopt a 'never appropriate for ad buys' position for media. The list below identifies the categories as well as how Oracle Contextual Intelligence is meeting these standards.

Mapping of ODC's Contextual Intelligence avoidance categories to the 4A's Advertising Assurance Brand Safety Floor Framework		
Category	4As Floor	ODC Category/Response
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	gv_adult category covers
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	gv_arms category covers
Crime & Harmful acts to individuals and Society, Human Right Violations	Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity – Explicit violations/demeaning offenses of Human Rights (e.g. human trafficking, slavery, self harm, animal cruelty etc.), Harassment or bullying of individuals and groups	gv_crime category covers

Death, Injury or Military Conflict	Promotion, incitement or advocacy of violence, death or injury Murder or Willful bodily harm to others Graphic depictions of willful harm to others Incendiary content provoking, enticing, or evoking military aggression Live action footage/photos of military actions & genocide or other war crimes	gv_death-injury category covers, plus gv_military. Current development ongoing to fuse these two.
Online piracy	Pirating, Copyright infringement, & Counterfeiting	gv_dowload category covers piracy-related content but cannot ascertain pirated content to date. Additionally as part of our monitoring, we manually add spam sites to our internal block-lists of pages not to be crawled. Included in this are criminal site lists provided by law enforcement agencies such as the London City Police Internet Crimes Unit (PIPCU)
Hate speech & acts of aggression	Behavior or content that incites hatred, promotes violence, vilifies, or dehumanizes groups or individuals based on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status, or serious disease sufferers.	gv_hate_speech category covers

Obscenity and Profanity, including Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult. language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.	gv_obscenity covers
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	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 Promotion and advocacy of Tobacco and e-cigarette (Vaping) & Alcohol use to minors	gv_drugs and gv_tobacco categories cover. current development ongoing to fuse the two.
Spam or Harmful Content	Malware/Phishing	As Contextual Intelligence can only identify content categories, the gv_download category can flag some. Additionally as part of our monitoring, we manually add spam sites to our internal block-lists of pages not to be crawled.
Terrorism	Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society	gv_terrorism covers

Debate Sensitive Social Issues	Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite greater conflict;	Due to the variable/subjective nature of this designation, there is no map directly to a specific ODC avoidance categories, although certain text related to this definition may be covered in some part by the “Hate speech” and “Obscenity” categories, and sites deemed inappropriate may be removed from our crawl list and be designated with a "harmful_site" categorization

In addition, the Brand Suitability framework advocates for additional layers beyond the safety floor to allow for advertisers to selectively avoid content according to the sensitivity of that brand. This 'tiered' standard attempts to separate standard content categories into 'low', 'medium', and 'high' risk categories. ODC is currently collaborating with the 4As around how to define these tiers, with an anticipation of deployment of a tiered solution in spring of 2021.

Maintaining URL Categorization Quality

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

1. First, an editorial categorization analyst requests at least 1,000 texts from engineering in the specified language.
2. A software engineer then provides a list of URLs and runs the categorization (gs_ and gv_) against the texts. The top three standard segment matches are noted for each text.
3. 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.[8]

[8] [1]https://en.wikipedia.org/wiki/Precision_and_recall

4. 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.[9]

Updates to the Crawler for Categorization

In addition to the processes described above, ODC takes action based on feedback from additional sources in order to maintain the high levels of 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

ODC Contextual Intelligence team maintains a database of domains on which content or crawler response behavior is likely to cause classification or brand safety issues. Domains on this list have either been detected automatically or are flagged by support and editorial staff.

Maintaining Segment Quality

To ensure the validity of ODC's standard contextual segments, we conduct regular quality assessments of them for each of the languages we cover. Our monitoring focuses on ensuring accuracy of contextual classification of brand safety segments, followed by contextual segment accuracy for targeting. There is a multi-step process for standard contextual segments available to all customers. 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 ODC-supervised segments.

The frequency of review is driven by size of the language footprint as well as the pace of change of topics. English, French, German, Japanese, Spanish, Italian and Chinese (both simplified and traditional) are typically reviewed w/ the greatest frequency. As the volume of syndicated segmentation grows, the frequency

[9] 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.

for repeated segment review is impacted. We are also working on automated monitoring solutions to flag segments for review when quality issues are detected.

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 ODC's technology and processes. Our teams 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 periodically check how many new terms we receive and accept so that we can reschedule monitoring frequency if we discover the need to do so. When we update the terms for inclusion, we test their validity by using the ODC bookmarklet against regional news articles. We also conduct periodic cross-team reviews to ensure consistency of our methods and processes.

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, all requests about 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. With consideration of all possible synchronization between ODC and client-side cache infrastructure, *on average*, changes will be surfaced in no more than **four hours,** although it can be faster.

Use of Metadata

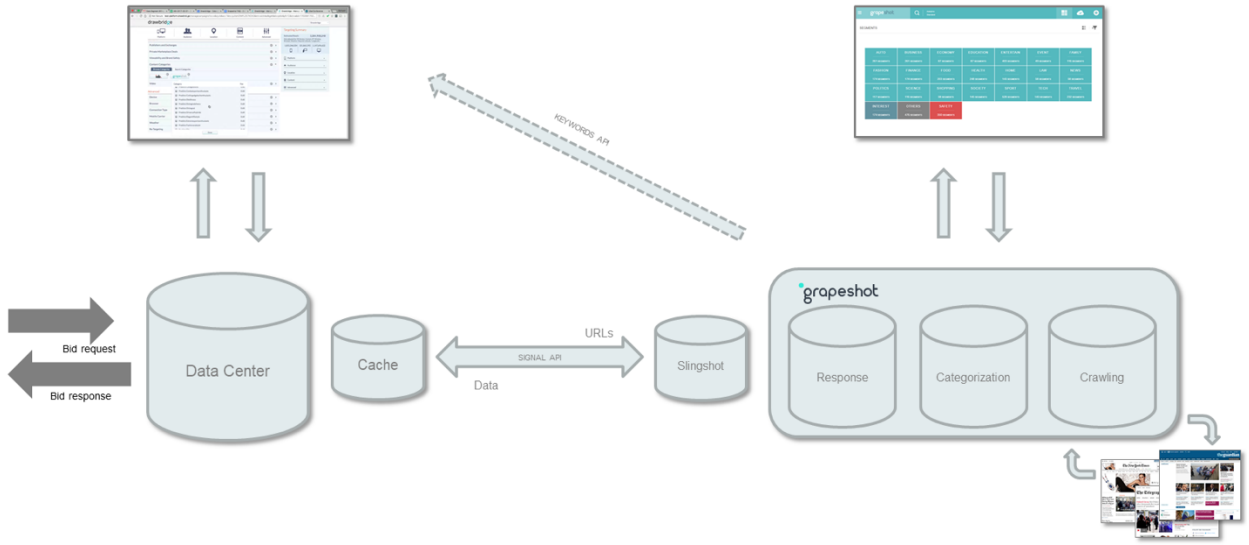
As described above, ODC's Contextual Intelligence engine relies most strongly on page text for categorization purposes, and may also include the “title,” "description" and “keywords” metadata fields as part of our consideration process. Generally, we do not rely on page tag data as we have found it to be highly variable and in some cases deliberately misleading. 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 ODC's Contextual Intelligence - The Methods

Integrations Overview

There are three primary ways for partners to interface with ODC'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 ODC's servers to give contextual analysis on page requests. The Keyword API is for handling and updating contextual segments.



As typical 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[10]: 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.

[10] ODC's Javascript tag when triggered on a page will execute Contextual Intelligence page categorization engine and return Contextual Intelligence page categorization results.