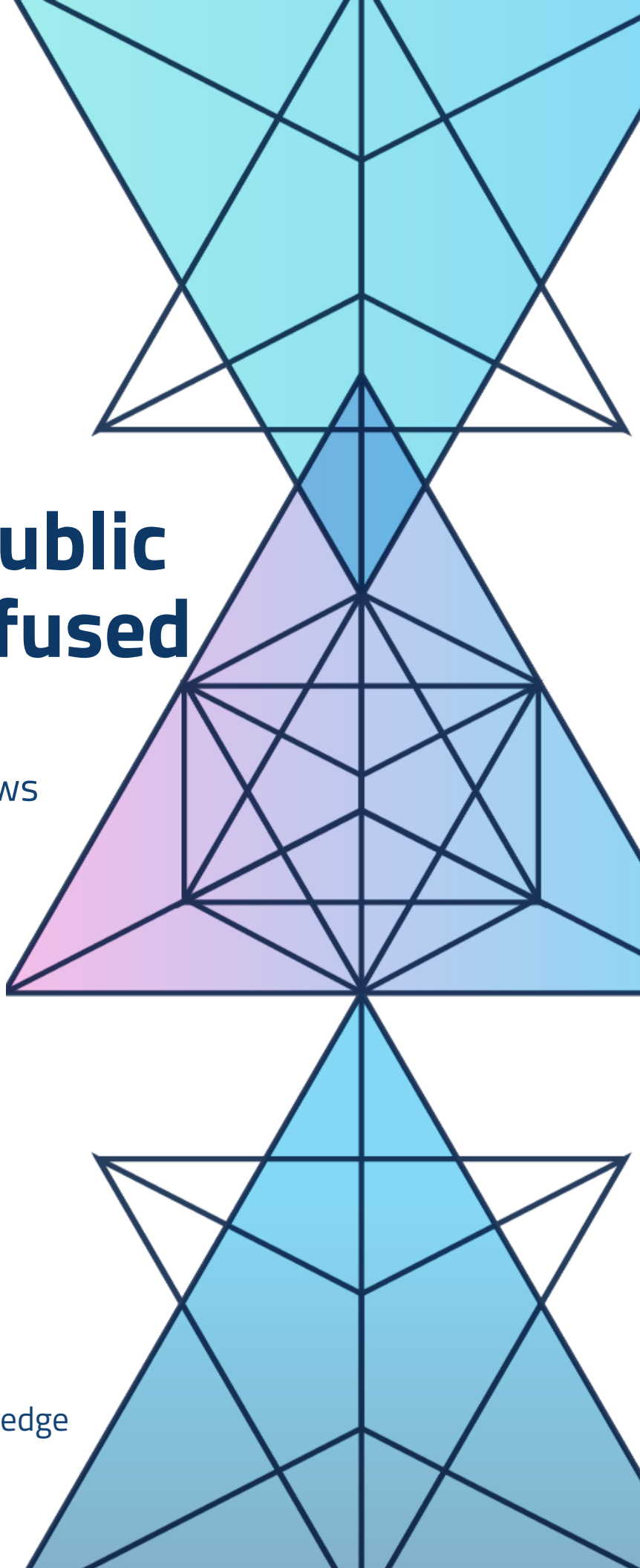
April 2021

Towards A Public Web Data Infused Dashboard

For market intelligence, news monitoring, and lead gen

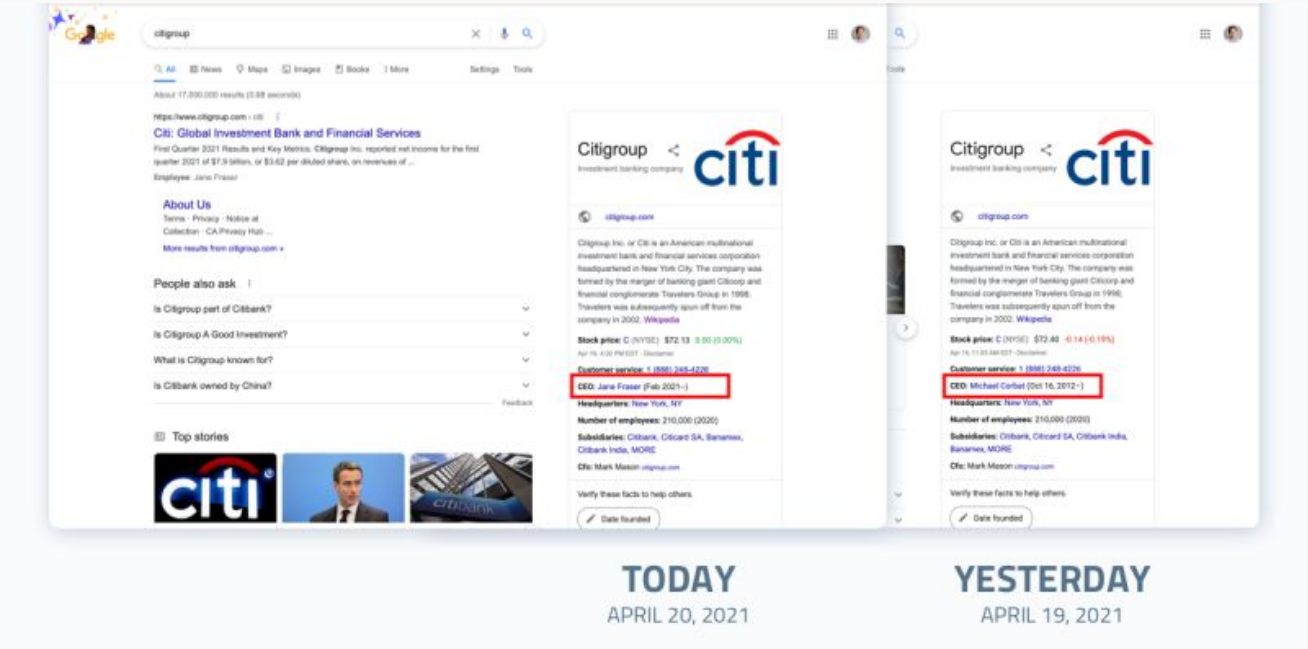


We structure
the world's knowledge

DIFFBOT

Intro

It took Google knowledge panels one month and twenty days to update following the inception of a new CEO at Citi, a F100 company. In Diffbot's Knowledge Graph, a new fact was logged within the week, with zero human intervention and sourced from the public web.



The CEO change at Citi was announced in September 2020, highlighting the reliance on manual updates to underlying Wiki entities.

In many studies data teams report spending 25-30% of their time cleaning, labelling, and gathering data sets [1]. While the number 80% is at times bandied about, an exact percentage will depend on the team and is to some degree moot. What we know for sure is that data teams and knowledge workers generally spend a noteworthy amount of their time procuring data points that are available on the

the public web.

The issues at play here are that the public web is our largest -- and overall -- most reliable source of many types of valuable information. This includes information on organizations, employees, news mentions, sentiment, products, and other “things.”

Simultaneously, large swaths of the web aren’t structured for business and analytical purposes. Of the few organizations that crawl and structure the web, most resulting products aren’t meant for anything more than casual consumption, and rely heavily on human input. Sure, there are millions of knowledge panel results. But without the full extent of underlying data (or skirting TOS), they just aren’t meant to be part of a data pipeline [2].

With that said, ***there’s still a world of valuable data on the public web.***

Online Footprints Accumulated in the Knowledge Graph

1.3B+ Article entities

715M+ Person entities

242M+ Organization entities

At Diffbot we’ve harnessed this public web data using web crawling, machine vision, and natural language understanding to build the world’s largest commercially-available Knowledge Graph. For more custom needs, we harness our automatic extraction APIs pointed at specific domains, or our natural language processing API in tandem with the KG.

DIFFBOT

In this paper we're going to share how organizations of all sizes are utilizing our structured public web data from a selection of sites of interest, entire web crawls, or in tandem with additional natural language processing to build impactful and insightful dashboards par excellence.

Note: you can replace "dashboard" here with any decision-enabling or trend-surfacing software. For many this takes place in a dashboard. But that's really just a visual representation of what can occur in a spreadsheet, or a Python notebook, or even a printed report.

DIFFBOT

Overview

The three dashboard-enhancing use cases we see the most regularly at Diffbot include use cases for market intelligence, news monitoring, and lead generation or enrichment. Additionally, when organizations can find a great deal of their most valuable data online, these methods can be combined in powerful ways we'll tackle in the final section of this paper.

DIFFBOT

Structured Web Data For Market Intelligence

Market intelligence has existed from the earliest days of commerce. Marketplaces of Ancient Rome, Mesopotamia, or early Chinese dynasties all had suppliers, competitors, salespeople, customers, “word on the street,” and products.

Today the main differences are three-fold:

- A vast majority of market participants have a trackable footprint (public web data)
- Markets are global
- Every market influencing action has more data attached

The public web is at the center of each of these differences. And while the bulk of (particularly informal) market intelligence is done through search engines, there is a caveat: **manually tracking down market intel data doesn’t scale.**

Market intel data tracked in stone doesn’t scale well either

DIFFBOT

So how do knowledge teams go about scaling up their market intelligence efforts?

Many of the most important data fields for market intel are publicly available online, but spread across the web and impractical to wrangle beyond a certain point.

Historically, automation of market intelligence data would depend on the type of data a team was after. For product data, teams might build their own web scraper to track prices on ecommerce sites. For news data, they could pipe RSS feeds to a dashboard. Data can also be purchased in bulk or subscription access for given domains.

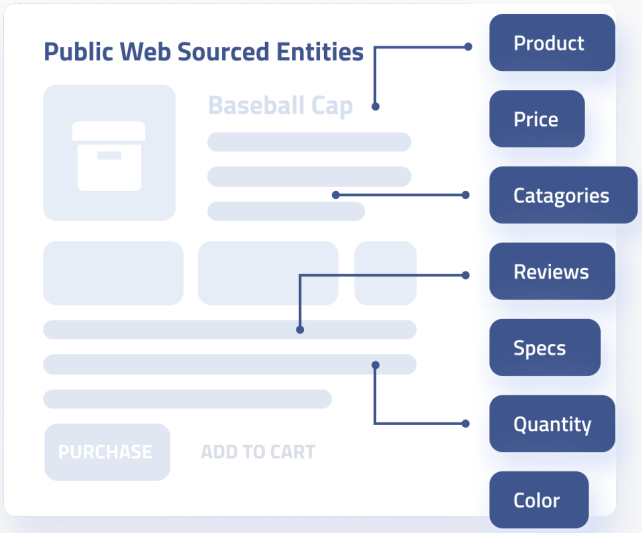
This piecemeal approach does enable some leveraging of web data, but involves maintaining an array of infrastructure types and talent.

Diffbot helps to streamline leveraging many of the most important market intelligence data types by organizing public web data into structured entities.

Some characteristics of these entities include:

- They represent people, places, or “things”... essentially the entities that matter within market research
- They’re linked, for example an organization entity is attached to article entities that reference it
- Facts or relationships attached to an entity have explicit sources for data provenance

DIFFBOT



Diffbot's market intel entities are sourced from the public web using NLP, machine vision, and web data extraction

Diffbot's entities correspond with prevalent topical data clusters within marketing intelligence:

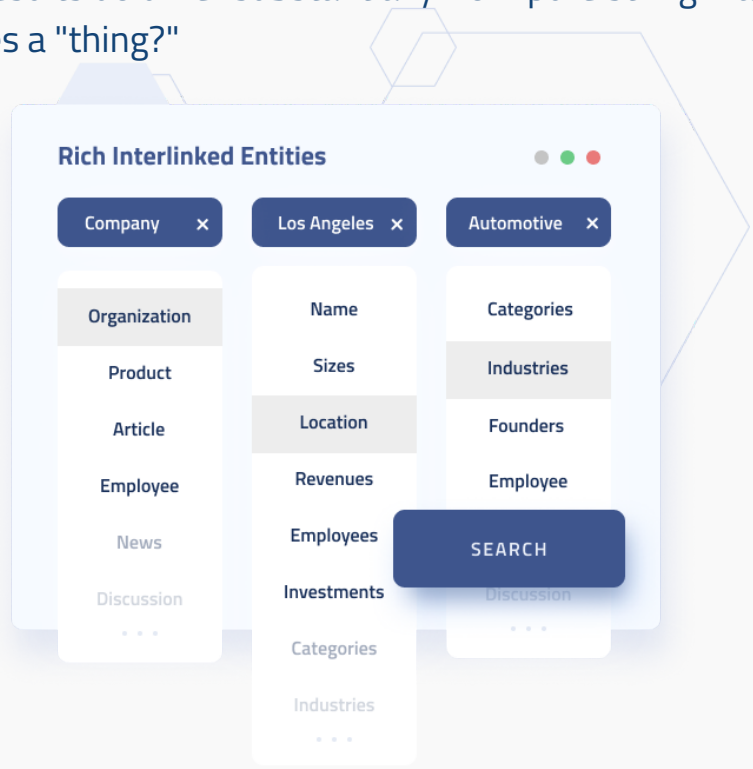
- **Organizational entities** provide firmographic data including industry codes, locations, funding rounds, employee counts, subsidiaries, key individuals, and news mentions.
- **Person entities** provide data on key individuals within organizations, hiring trends, and macro level views about where talent works and lives.
- **Product entities** provide data on products including fields like availability, SKUs, product options, review data, fraudulent products, pricing (including sales or internationalization of pricing), and more.
- **Article entities** provide news data on any of the above.
- **Discussion entities** provide discussion or review data on any of the above.
- **Event entities** provide data on past, present, and future events.

DIFFBOT

At a high level this aligns with the types of data that inform market intel inquiries. But how does it integrate with your workflow?

The difference between search engine search and knowledge graph search

Google's marketing department was the entity who coined "things not strings." And their search results do differ substantially from pure string matching. But what really differentiates a "thing?"



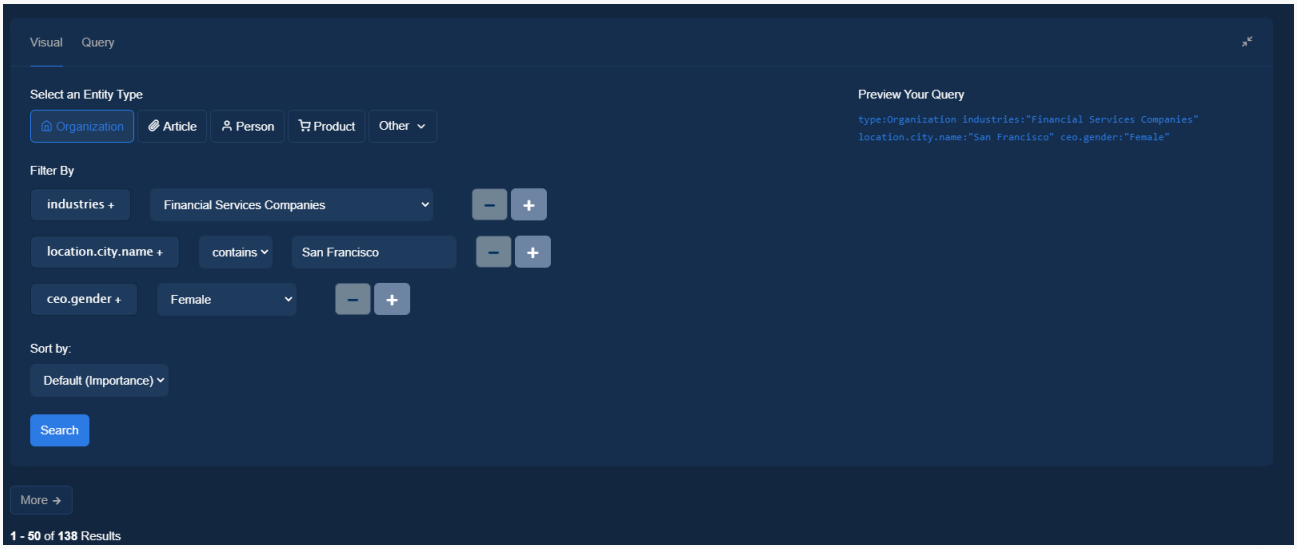
'Things' have properties specific to their 'type' useful for filtering

"Things" (aka entities) have different properties. Cars have makes, models, and colors. People have educational and employment histories, posts online, and skill sets. And organizations can have funding rounds, news mentions, employees, and industry codes.

The types of properties each entity can have are described in a **taxonomy**. Each entity type has facts that pertain to it. Facts that matter about an article are different than facts about a startup.

Search engines do track a great many of these entities and properties. Many are populated on knowledge panels. But searchers are still at the mercy of the search engine as to whether it deems their search worthy of showing a knowledge panel. There's no summary view or public repository of a list of knowledge panel entities. Even if you compiled one, there's no export or API access. **This entity-centered data isn't built or productized for inclusion in your data pipeline.**

This points to perhaps the most prescient difference between commercially available knowledge graphs and search engine results. They're both forms of



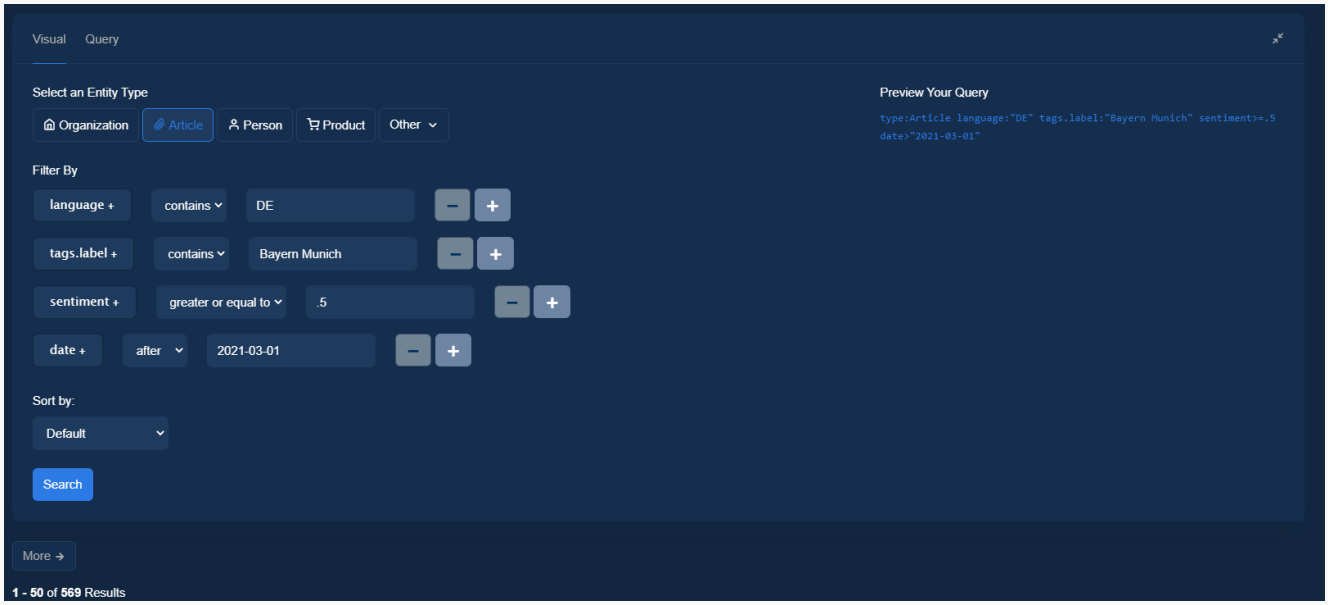
With KG search you filter through entities. In this case we can see all finance firms based in San Francisco with female CEOs (138 results)

DIFFBOT

"search," but only one lets you interact with the underlying entity and relationship data.

So what does KG search look like?

In Diffbot's Knowledge Graph you can traverse, sort, and facet billions of entities by their properties or facts about them.



In this query we filter article entities to find all very positive sentiment German-language articles about soccer team Bayern Munich from the last month

In this query we seek out a summary view of skills among all of Microsoft's employees. Facet searches like above can be great for market intel.

And you can export this data, either via API to a dashboard or product, in integrations, or by downloading in CSV or JSON. All mentioned entities with all fact fields are included.

This is search beyond text-based content using the world's largest commercially-available database of firmographic, demographic, and article-centered facts.

Web Wide or Targeted Public Web Data

Thus far we've talked about the types of data our Knowledge Graph can provide. And the Knowledge Graph alone does power many dashboards. But as with any entity that crawls the entire web, there are far corners that may not get visited with enough regularity for your needs. In this case, a different method for structuring specific locations online is needed.

DIFFBOT

And this comes in the form of web scraping.

Diffbot's Automatic APIs are built off of the same underlying tech that enables the Knowledge Graph. They can handle a vast majority of the most common page types online without any rules required. *In fact, they're built to work when being sent to pages we don't know the structure of in advance.*

Depending on the type of entity you want to monitor, you'll need to select an Automatic API that fits. For example, the Product API can be pointed at product pages to extraction a wide range of fields including reviews. The Article API can extract articles, topical tags, and sentiment, and be paired directly with Diffbot's Natural Language API.

Leveraging crawls that you point to a destination has the benefit of enabling you to control how often to refresh your market intelligence.

Additionally, for pages then parsed with our Natural Language API, data can be returned in a graph structure that can be integrated directly into our larger Knowledge Graph.

Takeaways

- Market intel inquiries are better suited with "things" not "strings."
- Linked data structures like graphs can surface relationships between people, products, articles, and organizations.
- With breadth of data coverage comes some lag time. For close to "live" public web data, you may need to schedule your own crawls of specific domains.

DIFFBOT

Structured Web Data For News Monitoring

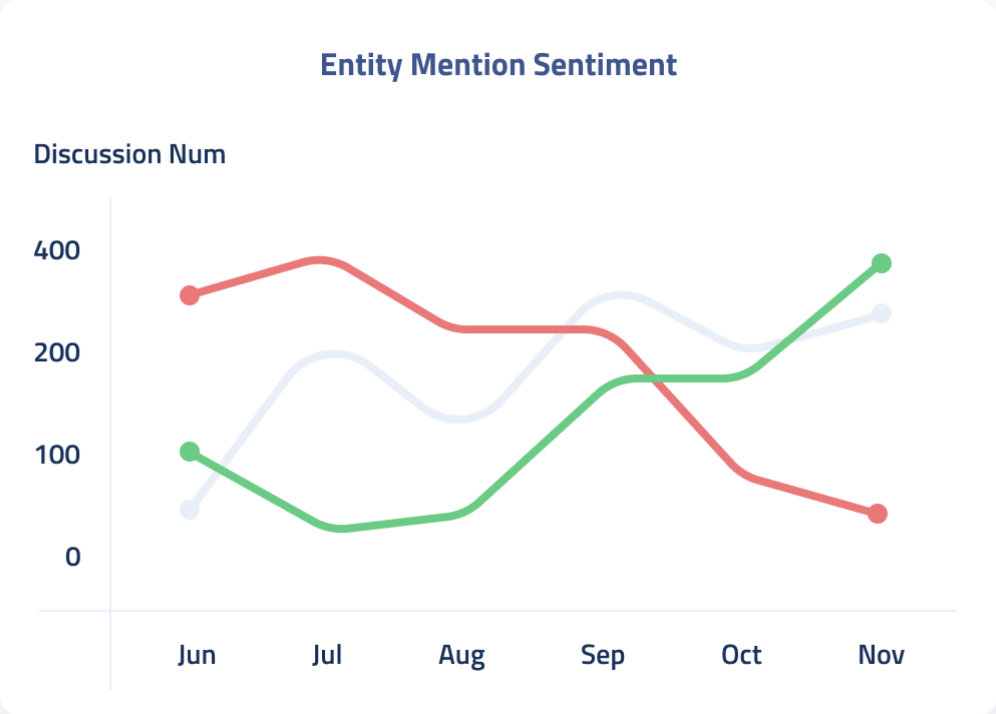
For several decades now, each passing year has seen exponentially increasing news data volume. When you factor in informal news such as “mentions,” reviews, or social posts, the bulk of news-related data grows yet again.

In 2021 the public web is by far the largest source of this news. And while there are a range of ways to integrate news data into a dashboard, we’ve heard many client stories about what works and what doesn’t.

RSS feeds and the use of social media tracking tools have historically been the two most prevalent forms of news monitoring tracking. But these methods require you to know where your news of interest is coming from. Manual research is also still a norm, though for both topical clusters occupying particularly long tail locations online, or that are widespread, scaling this form of news monitoring is impractical.

A few rules of thumb about what makes valuable news monitoring data in many dashboards we have seen include:

- News data that is both structured and has some additional processing is most valuable
- News data that can flag entities is valuable
- Determine if you want to monitor a selection of sites or individuals more regularly
- Pass article data through NLP to derive additional fields



Many of the 1B+ articles in the KG have document-level sentiment. By passing each document to our NL API teams derive entity-level sentiment.

News data that is both structured and has some additional processing is valuable. For a dashboard to be useful, presented entities need enough fields to be able to sort, pull out multiple trend types, and be fully accessible to mine into. Common fields we see of interest include article or review sentiment, topical tags, mentions of entities, speakers of quotes, and the ability to process coverage in many languages. All of these field types are available in our Knowledge Graph article index as well as our Article Extraction API results.

For the ability to add additional fields including entity level sentiment and salience, facts and relationships between entities, and a graph structure to your results, our

DIFFBOT

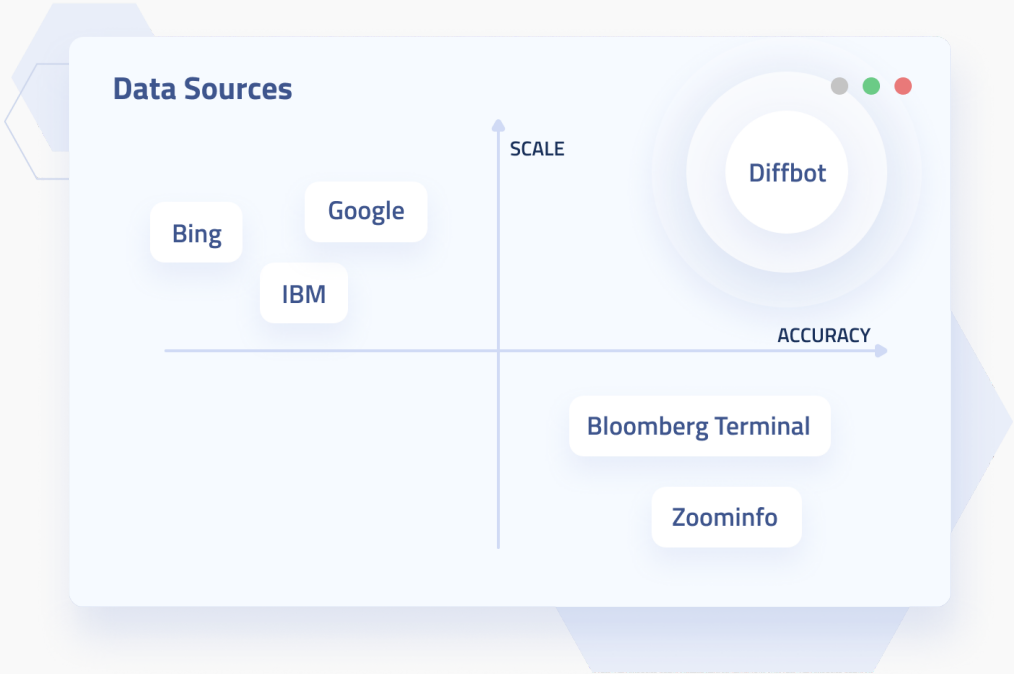
Natural Language API can help.

Granular entity-level sentiment and salience in NL API results

News data that can flag entities is valuable. Rather than simply pulling in content that centers around a topic or was published by a specific site, enable dashboard users to filter by a preset list of topical tags. In conjunction with sentiment data, this can give you a more granular view of what's going right and wrong in public discourse about an entity online.

Topical tags come standard in articles within our Knowledge Graph news index

(roughly 50x the size of the Google News Index) and from articles parsed with our Article Extraction API. Our Natural Language API can provide additional fields including how central a topic is to understanding text, and entity level sentiment.



Web-wide crawls that power the KG take longer to refresh, but can surface entities from the far-flung corners of the internet.

Determine if you want to monitor a selection of sites or individuals more regularly. We routinely see wide media monitoring coupled with selective media monitoring to great effect. Cast a wide net on subjects you don't know what type of coverage will occur about. Then for more predictable sources of mentions like review sites or social media, keep tabs on individual pages or specific domains.

Pass article data through NLP to derive additional fields. Natural language processing can provide unmatched ability to understand -- particularly technical or informal language -- at scale. Diffbot's NL API can be passed an article or review

text directly from our Knowledge Graph or extraction APIs to provide entities, facts and relationships between entities, sentiment and salience of entities, and more.

Additionally, Diffbot's NL API enables you to train models based on entities and fact types you care about. Train recognition of up to 1M custom entities in as little as a day to gain a fully-tailored NLP dashboard feed, or to continue tuning as your needs change.

Takeaways

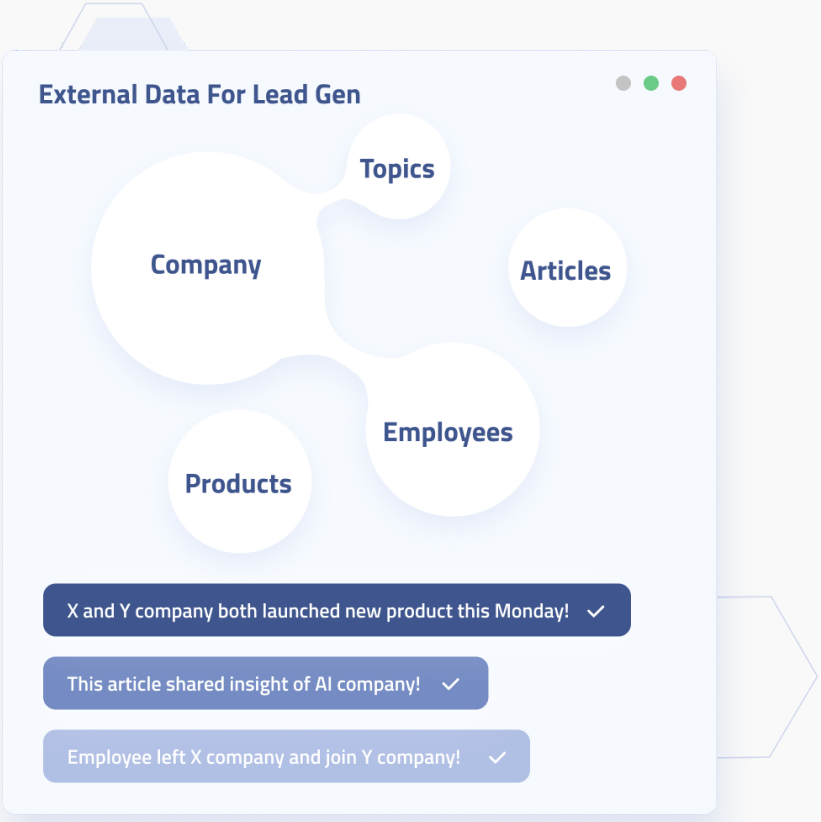
- Based on the scope of news monitoring efforts, you may be able to utilize an API, RSS feed, use web scraping, or need to utilize a service that crawls the web.
- Some news monitoring services can provide multilingual content, while others deal primarily in one language.
- Determine whether you need raw article or mention data or an additional level of processing such as provided by natural language processing services.
- Determine the regularity with which you need updated news
- For full web monitoring services like Diffbot's Knowledge Graph can provide global coverage that's updated every few days
- For coverage of specific sites with greater regularity setting up web scrapers that run on your schedule may be the right option

DIFFBOT

Structured Web Data For Lead Generation And Enrichment

More than 6 in 10 marketers rank lead generation as their largest challenge [3]. And over 1 in 2 marketers spend over half of their budget on lead generation [4]. And yet the basics components of lead generation are surprisingly simple.

Sure, there are different methods. But market segmentation, creation of an ideal customer profile, and accumulation of outreach lists are fairly standard.



Interlinked data facilitates locating orgs of interest, then mining into people, products, mentions, and events

In most domains this process involves accumulating firmographic data and then mining into which roles and specific individuals you want to reach out to at what time, and about what topic. In many cases, all of this information is publicly available, but spread across the web, and is a manual process where you may need to perform repetitive research for each entity you're appraising.

By pulling from linked firmographic data that is pre-structured from across the web a number of lead gen benefits actualize:

- Iterate through ICP hypotheses faster
- Discover "lookalike" entities
- Reach much more long tail entities
- Personalize messaging and outreach timing
- Augment your existing data on entities
- Data that's "equitable" or ethically sourced

In Diffbot's Knowledge Graph, the specific fields most useful for lead generation may include:

- Many fields attached to organization entities including industry code, employee skills, location, subsidiaries, sentiment of news mentions
- The linked quality of KG data which enables users to mine into person data (skills, locations, news mentions) attached to specific organizations

Lead generation is, after all, about reaching people with the backdrop that they're related to an organization of interest.

Next we'll look at a few unique characteristics of Diffbot's data when applied to lead generation.

DIFFBOT

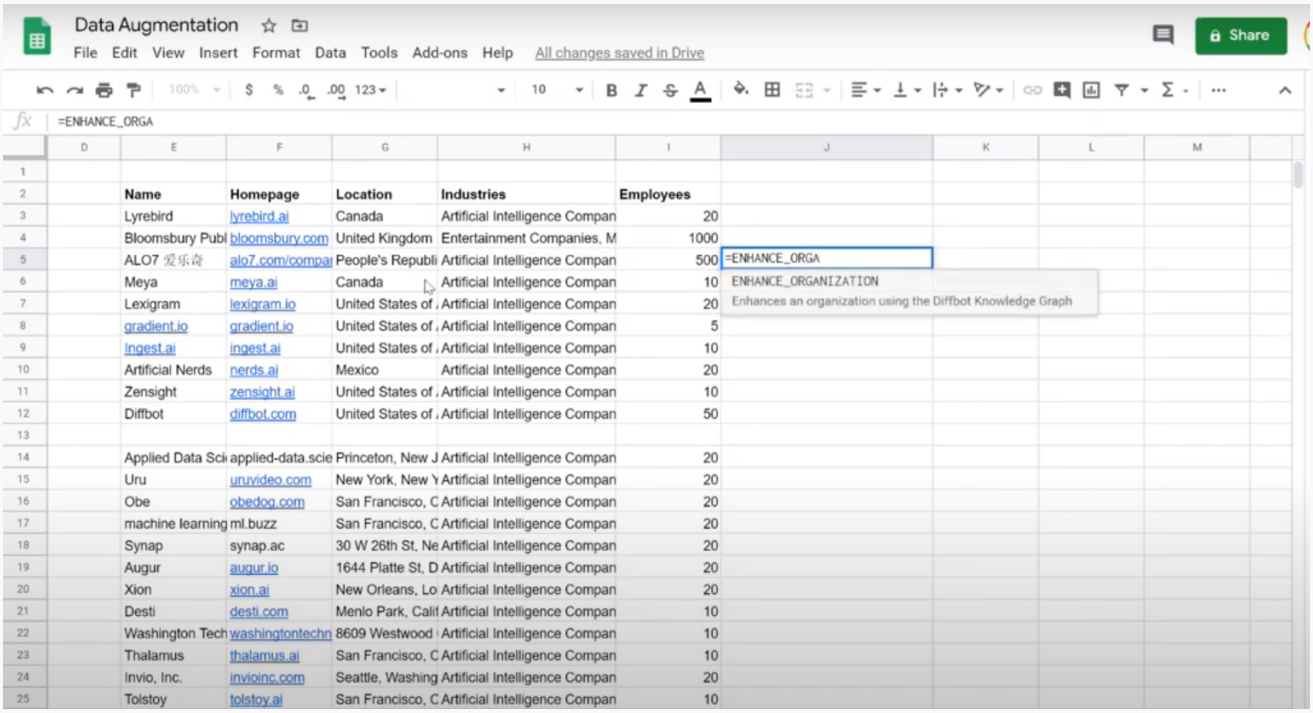
Equitable Web Data Sourcing
At times we call it “guilt-free” data.

It’s the data that’s been willingly shared online or inferred from mentions in reputable news or reporting sites. It’s not the data that’s bought and sold from data breaches or from “that form you filled out to get a white paper.”

This is unique within “lead gen,” where inMail and email boxes are filled daily with solicitations about “shady” data.

And while at the end of the day you will need a way to contact individuals who are part of outbound lead gen, you don’t have to deal in this data for the parts of the process that provide the bedrock for lead gen success. Namely, insight into your ideal customers, larger org ecosystems, and news mentions from across the web.

And that’s where our equitable public web data comes in. It’s not data that’s been bought. It’s data that’s been abstracted and structured from our collective coverage (with millions of contributors): the public web as a whole.



	Name	Homepage	Location	Industries	Employees
3	Lyrebird	lyrebird.ai	Canada	Artificial Intelligence Compan	20
4	Bloomsbury Publ	bloomsbury.com	United Kingdom	Entertainment Companies, M	1000
5	ALO7 爱乐奇	alo7.com/compai	People's Republi	Artificial Intelligence Compan	500
6	Meya	meya.ai	Canada	Artificial Intelligence Compan	10
7	Lexigram	lexigram.io	United States of	Artificial Intelligence Compan	20
8	gradient.io	gradient.io	United States of	Artificial Intelligence Compan	5
9	Ingest.ai	ingest.ai	United States of	Artificial Intelligence Compan	10
10	Artificial Nerds	nerds.ai	Mexico	Artificial Intelligence Compan	20
11	Zensight	zensight.ai	United States of	Artificial Intelligence Compan	10
12	Diffbot	diffbot.com	United States of	Artificial Intelligence Compan	50
14	Applied Data Sci	applied-data.scie	Princeton, New J	Artificial Intelligence Compan	20
15	Uru	uruvideo.com	New York, New Y	Artificial Intelligence Compan	20
16	Obe	obedog.com	San Francisco, C	Artificial Intelligence Compan	20
17	machine learning	ml.buzz	San Francisco, C	Artificial Intelligence Compan	20
18	Synap	synap.ac	30 W 26th St, Ne	Artificial Intelligence Compan	20
19	Augur	augur.io	1644 Platte St, D	Artificial Intelligence Compan	20
20	Xion	xion.ai	New Orleans, Lo	Artificial Intelligence Compan	20
21	Desti	desti.com	Menlo Park, Calif	Artificial Intelligence Compan	10
22	Washington Tech	washingtontechn	8609 Westwood	Artificial Intelligence Compan	10
23	Thalamus	thalamus.ai	San Francisco, C	Artificial Intelligence Compan	10
24	Invio, Inc.	invioinc.com	Seattle, Washing	Artificial Intelligence Compan	20
25	Tolstoy	tolstoy.ai	San Francisco, C	Artificial Intelligence Compan	10

Pull the KG into spreadsheets for a quick exploratory analysis or an informal dashboard. Enhance enriches data you already know something about.

Data That Lives Where You Work

While we've mentioned that most valuable lead generation data is available publicly and free of charge, it's not efficiently actionable until it's living where you work.

Diffbot's Enhance product provides data enrichment from within (y)our dashboard or directly from other programs like Excel. Enhance pores over the Knowledge Graph like KG search, but uses a different matching algorithm that seeks to return one match for an organization or person you hold incomplete data about.

Partial, flawed, duplicate, or under merged entities can be corrected with KG data via Enhance.

CRM Cleanups

The average company loses 12% of revenue due to bad data [4]. And if there's one data store at ground zero for revenue for most companies, it's the CRM.

Luckily for a database like Diffbot's Knowledge Graph, even the ugliest CRM record likely has at least one field of overlap with the trillion+ facts in the KG. Our Enhance product can take lists of partial, un-de-duplicated, and incorrect organizational or person data and return a rich set of fields from across the web.

More than a CRM cleanup, we see Enhance used daily to augment the fields that make CRMs work. The inclusion of social profiles, news mentions around organizations, industry codes, business sizes, and more lead to personalization, filtering out leads who aren't actually a good fit, and overall better sales outcomes.

DIFFBOT

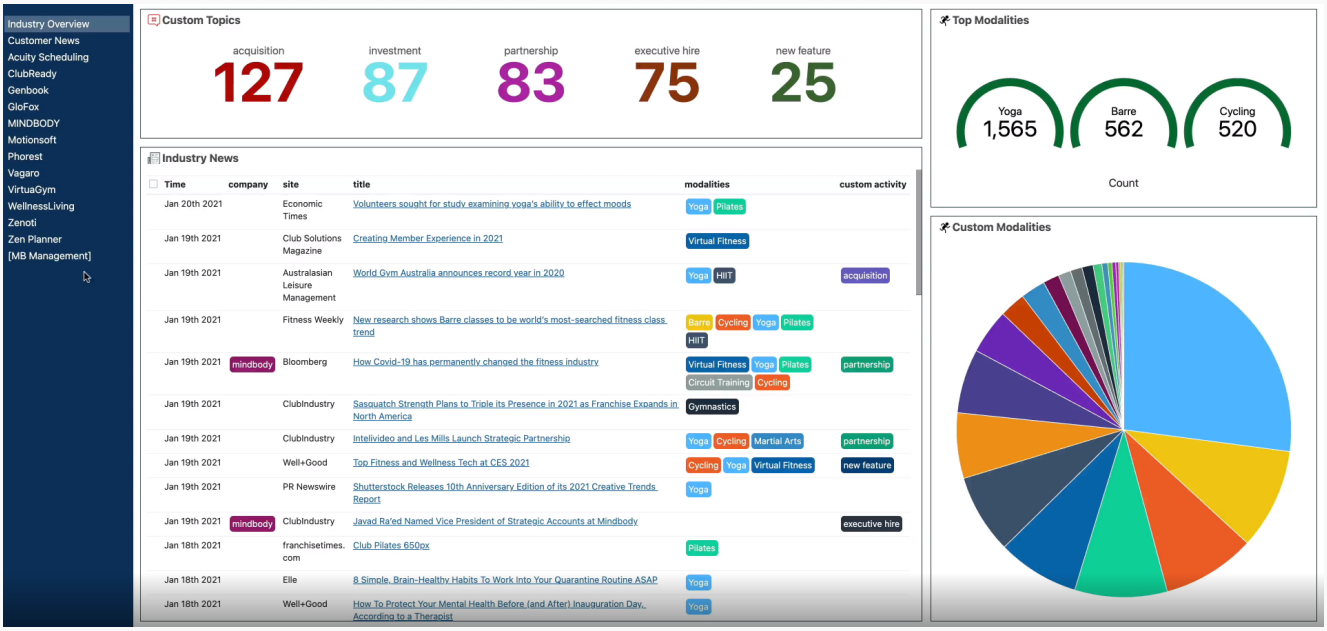
Pulling It All Together

If you've made it this far, you've seen how linked public web data can serve as the bedrock for multiple business functions. While these may live in separate dashboards, the use of one standardizing entity set leads to a number of benefits:

- A unique entity identifier that helps with over or under merging and deduplication
- The ability to augment market intelligence, news monitoring, or lead generation data as new data surfaces from across the web
- The ability to extend exploratory research beyond dashboards within the Knowledge Graph or in other BI programs

While the "right" dashboard for your team will depend on a ton of factors unique to you, we have some examples of high impact use of web data in dashboards. Below we've mocked up a representation of some of the ways structured public web data is best utilized within dashboards.

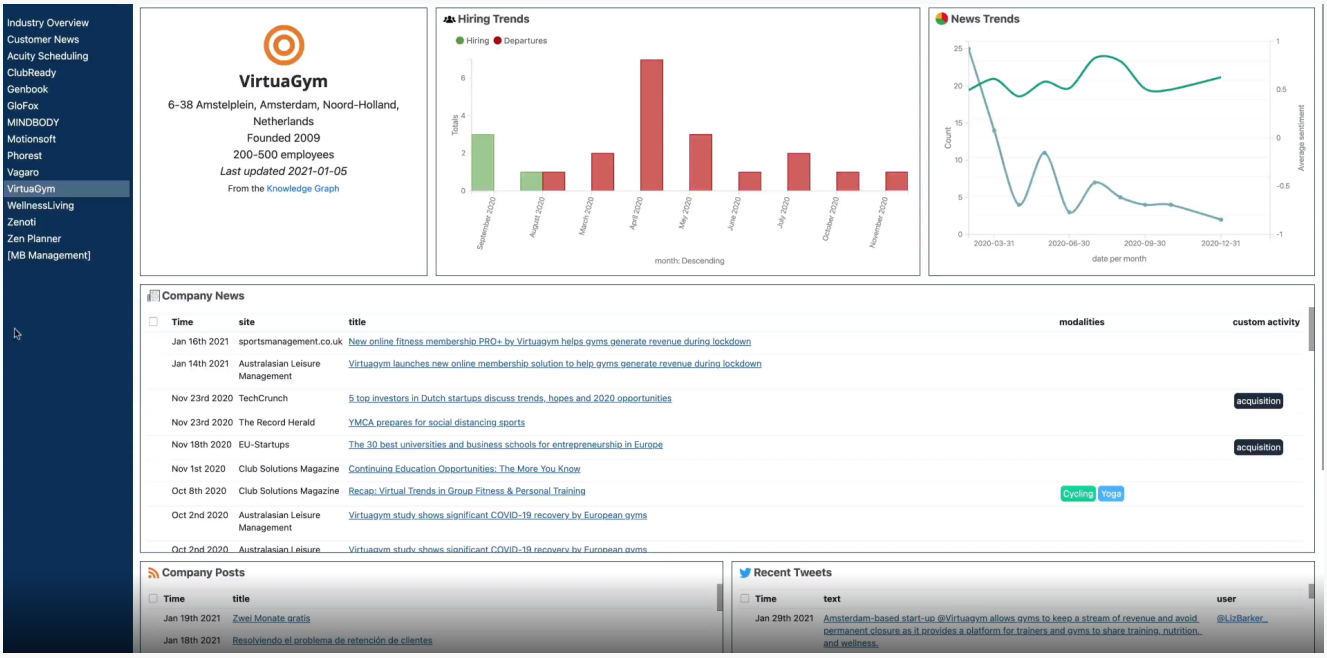
DIFFBOT



A custom dashboard that utilizes custom crawl data, the Knowledge Graph, and the NL API to explore aggregate values

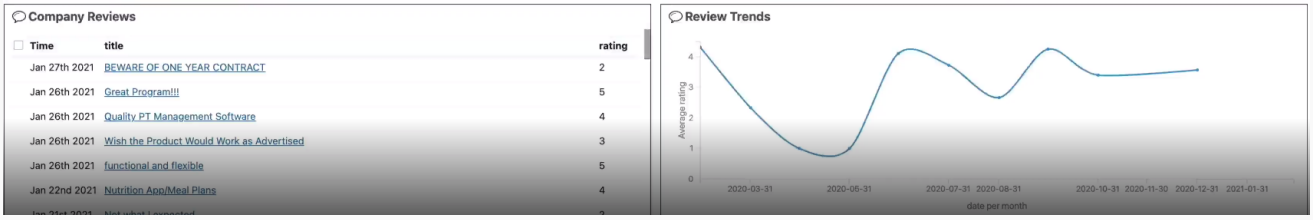
In our first dashboard, news monitoring data further processed through our Natural Language API is combined for a range of valuable aggregate values.

- "Top Modalities" serves to show velocity of publishing within a topical cluster
- "Custom Topics" shows where our NL API has identified specific events
- "Industry News" shows an overview of an entire industry sortable by a range of values and with all article data attached in event a full read is warranted
- With all underlying data included, you can always mine in to individual articles or entities that catch your eye.

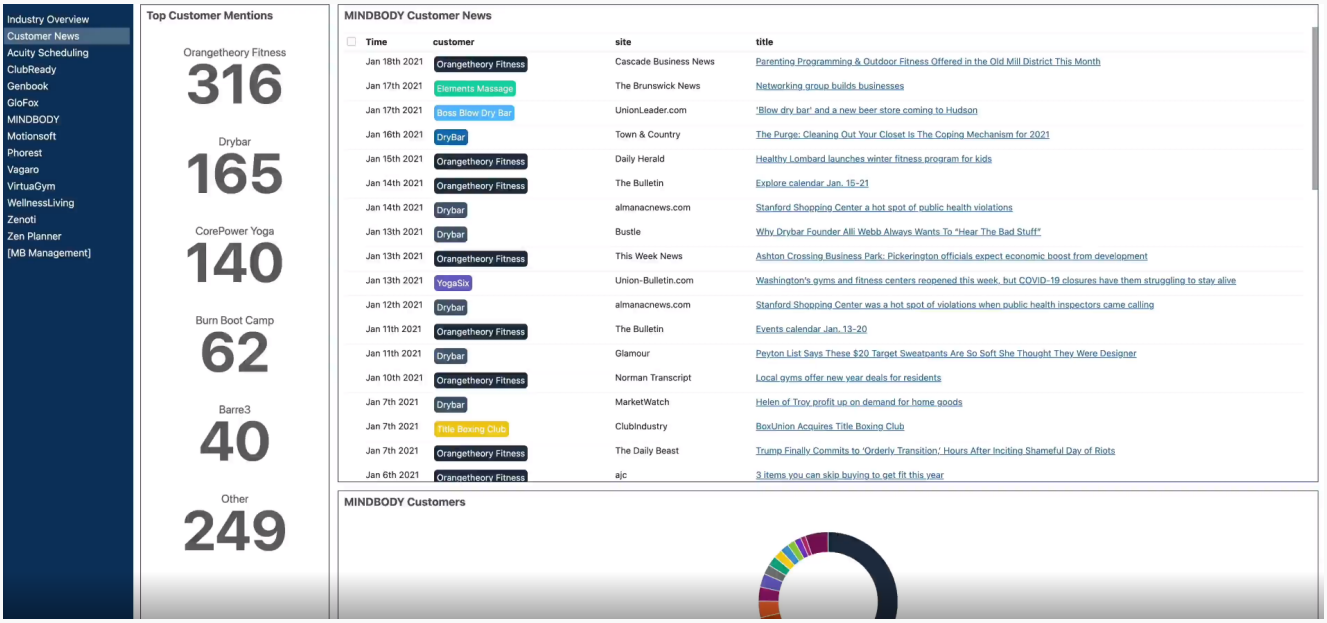


Firmographic and news monitoring data becomes exceedingly hard to track manually. Pair both for comprehensive coverage of an entity online

In our second dashboard view, news monitoring pairs with firmographic data on a competitor. Hiring trends are pitted against news mention velocity and sentiment.



Additionally, a view of external news mentions alongside company specific posts gives a particular granular view into how a company is handling a downsizing. In tandem with review data showing how this entity's products are actually perceived gives a well-rounded vision of our hypothetical competitor.



Each collection of articles tagged as topical and each customer organization can be entered with robust paired data from KG, NLP, and crawling

In our third dashboard example, we see the power of being able to list all of the data points listed in the past two dashboards for a range of competitors. Note that even in this example with less than 20 competitors, we've already outpaced the news monitoring ability of all but the largest teams performing manual monitoring.

Takeaways

Our example above is just a single manifestation of public web data for some of the purposes listed in our guide. While there are domains for which public web data is useful to greater or lesser extent, external data tends to be the least well utilized data source for may organizations (because it's unstructured). It should also be noted that manual research (eg how market research and news monitoring is often done) can't scale even to a medium-sized organization's needs.

A few takeaways about public web data fed to your dashboard include:

- The public web isn't built for data pipelines. It has to be structured first.
- Most organizations who crawl and structure the web don't release data on relationships between entities
- Long tail coverage for full web crawls is unrivaled in many domains
- Linked data allows for a connection between the "things" that matter in market intelligence and lead generation (orgs, people, products, articles)
- Market research hasn't fundamentally changed in a long time. Rather, more market participants have trackable footprints online.
- Public web data doesn't have to be unstructured, if you let us structure it for you!

Want to supercharge your market intel, news monitoring or lead gen dashboard?

Reach out at sales@diffbot.com or sign up for a free 14-day trial.

DIFFBOT

Citations

1. https://blog.ldodds.com/2020/01/31/do-data-scientists-spend-80-of-their-time-cleaning-data-turns-out-no/
2. https://stackoverflow.com/questions/22657548/is-it-ok-to-scrape-data-from-google-results
3. https://www.hubspot.com/marketing-statistics
4. https://televerde.com/how-not-to-blow-half-your-marketing-budget-on-lead-generation/
5. https://www.datamentors.com/blog/how-much-dirty-data-costing-you