

We've crawled the web for 32 years: What's changed?

It was 20 years ago this year that I authored a book called "Search Engine Marketing: The Essential Best Practice Guide." It is generally regarded as the first comprehensive guide to SEO and the underlying science of information retrieval (IR).

I thought it would be useful to look at what I wrote back in 2002 to see how it stacks up today. We'll start with the fundamental aspects of what's involved with crawling the web.

It's important to understand the history and background of the internet and search to understand where we are today and what's next. And let me tell you, there is a lot of ground to cover.

Our industry is now hurtling into another new iteration of the internet. We'll start by reviewing the groundwork I covered in 2002. Then we'll explore the present, with an eye toward the future of SEO, looking at a few important examples (e.g., structured data, cloud computing, IoT, edge computing, 5G),

All of this is a mega leap from where the internet all began.

Join me, won't you, as we meander down search engine optimization memory lane.

An important history lesson

We use the terms world wide web and internet interchangeably. However, they are not the same thing.

You'd be surprised how many don't understand the difference.

The first iteration of the internet was invented in 1966. A further iteration that brought it closer to what we know now

was invented in 1973 by scientist Vint Cerf (currently chief internet evangelist for Google).

The world wide web was invented by British scientist Tim Berners-Lee (now Sir) in the late 1980s.

Interestingly, most people have the notion that he spent something equivalent to a lifetime of scientific research and experimentation before his invention was launched. But that's not the case at all. Berners-Lee invented the world wide web during his lunch hour one day in 1989 while enjoying a ham sandwich in the staff café at the CERN Laboratory in Switzerland.

And to add a little clarity to the headline of this article, from the following year (1990) the web has been crawled one way or another by one bot or another to this present day (hence 32 years of crawling the web).

Why you need to know all of this

The web was never meant to do what we've now come to expect from it (and those expectations are constantly becoming greater).

Berners-Lee originally conceived and developed the web to meet the demand for automated information-sharing between scientists in universities and institutes around the world.

So, a lot of what we're trying to make the web do is alien to the inventor and the browser (which Berners-Lee also invented).

And this is very relevant to the major challenges of scalability search engines have in trying to harvest content to index and keep fresh, at the same time as trying to discover and index new content.

Search engines can't access the entire web

Clearly, the world wide web came with inherent challenges. And that brings me to another hugely important fact to highlight.

It's the "pervasive myth" that began when Google first launched and seems to be as pervasive now as it was back then. And that's the belief people have that Google has access to the entire web.

Nope. Not true. In fact, nowhere near it.

When Google first started crawling the web in 1998, its index was around 25 million unique URLs. Ten years later, in 2008, they announced they had hit the major milestone of having had sight of 1 trillion unique URLs on the web.

More recently, I've seen numbers suggesting Google is aware of some 50 trillion URLs. But here's the big difference we SEOs all need to know:

- Being aware of some 50 trillion URLs does **not** mean they are all crawled and indexed.

And 50 trillion is a whole lot of URLs. But this is only a tiny fraction of the entire web.

Google (or any other search engine) can crawl an enormous amount of content on the surface of the web. But there's also a huge amount of content on the "deep web" that crawlers simply can't get access to. It's locked behind interfaces leading to colossal amounts of database content. As I highlighted in 2002, crawlers don't come equipped with a monitor and keyboard!

Also, the 50 trillion unique URLs figure is arbitrary. I have no idea what the real figure is at Google right now (and they have no idea themselves of how many pages there really are on

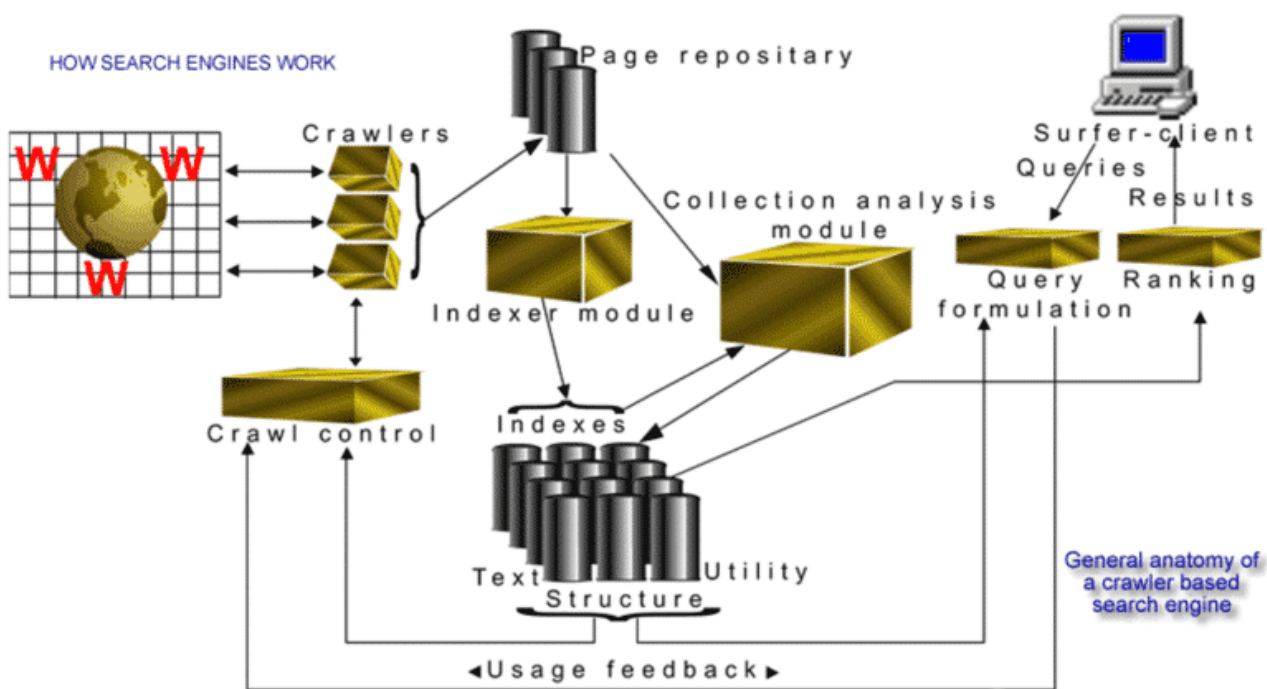
the world wide web either).

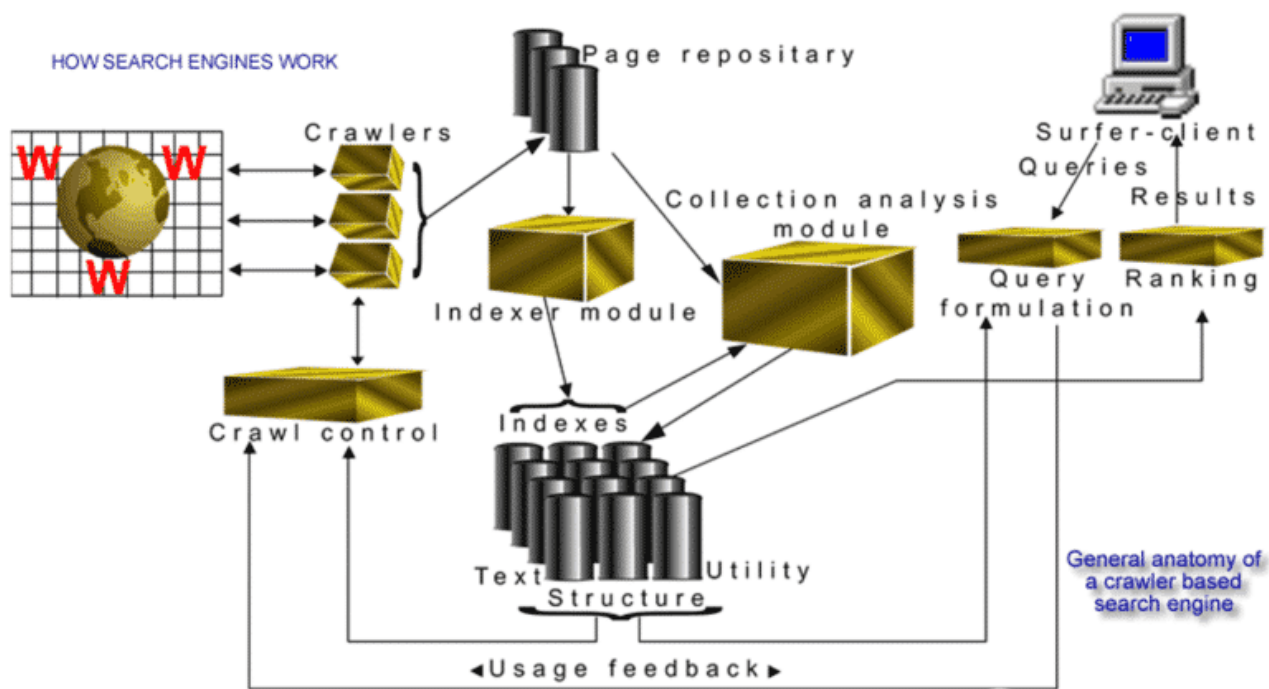
These URLs don't all lead to unique content, either. The web is full of spam, duplicate content, iterative links to nowhere and all sorts of other kinds of web debris.

- **What it all means:** Of the arbitrary 50 trillion URLs figure I'm using, which is itself a fraction of the web, only a fraction of that eventually gets included in Google's index (and other search engines) for retrieval.

Understanding search engine architecture

In 2002, I created a visual interpretation of the "general anatomy of a crawler-based search engine":





Clearly, this image didn't earn me any graphic design awards. But it was an accurate indication of how the various components of a web search engine came together in 2002. It certainly helped the emerging SEO industry gain a better insight into why the industry, and its practices, were so necessary.

Although the technologies search engines use have advanced greatly (think: artificial intelligence/machine learning), the principal drivers, processes and underlying science remain the same.

Although the terms “machine learning” and “artificial intelligence” have found their way more frequently into the industry lexicon in recent years, I wrote this in the section on the anatomy of a search engine 20 years ago:

"In the conclusion to this section I'll be touching on 'learning machines' (vector support machines) and artificial intelligence (AI) which is where the field of web search and retrieval inevitably has to go next."

‘New generation’ search engine crawlers

It’s hard to believe that there are literally only a handful of general-purpose search engines around the planet crawling the web, with Google (arguably) being the largest. I say that because back in 2002, there were dozens of search engines, with new startups almost every week.

As I frequently mix with much younger practitioners in the industry, I still find it kind of amusing that many don’t even realize that SEO existed before Google was around.

Although Google gets a lot of credit for the innovative way it approached web search, it learned a great deal from a guy named Brian Pinkerton. I was fortunate enough to interview Pinkerton (on more than one occasion).

He’s the inventor of the world’s first full-text retrieval search engine called WebCrawler. And although he was ahead of his time at the dawning of the search industry, he had a good laugh with me when he explained his first setup for a web search engine. It ran on a single 486 machine with 800MB of disk and 128MB memory and a single crawler downloading and storing pages from only 6,000 websites!

Somewhat different from what I wrote about Google in 2002 as a “new generation” search engine crawling the web.

“The word ‘crawler’ is almost always used in the singular; however, most search engines actually have a number of crawlers with a ‘fleet’ of agents carrying out the work on a massive scale. For instance, Google, as a new generation search engine, started with four crawlers, each keeping open about three hundred connections. At peak speeds, they downloaded the information from over one hundred pages per second. Google (at the time of writing) now relies on 3,000 PCs running Linux, with more than ninety terabytes of disk

storage. They add thirty new machines per day to their server farm just to keep up with growth.”

And that scaling up and growth pattern at Google has continued at a pace since I wrote that. It’s been a while since I saw an accurate figure, but maybe a few years back, I saw an estimate that Google was crawling 20 billion pages a day. It’s likely even more than that now.

Hyperlink analysis and the crawling/indexing/whole-of-the-web conundrum

Is it possible to rank in the top 10 at Google if your page has never been crawled?

Improbable as it may seem in the asking, the answer is “yes.” And again, it’s something I touched on in 2002 in the book:

From time to time, Google will return a list, or even a single link to a document, which has not yet been crawled but with notification that the document only appears because the keywords appear in other documents with links, which point to it.

What’s that all about? How is this possible?

Hyperlink analysis. Yep, that’s backlinks!

There’s a difference between crawling, indexing and simply being aware of unique URLs. Here’s the further explanation I gave:

“If you go back to the enormous challenges outlined in the section on crawling the web, it’s plain to see that one should never assume, following a visit from a search engine

spider, that ALL the pages in your website have been indexed. I have clients with websites of varying degrees in number of pages. Some fifty, some 5,000 and in all honesty, I can say not one of them has every single page indexed by every major search engine. All the major search engines have URLs on the "frontier" of the crawl as it's known, i.e., crawler control will frequently have millions of URLs in the database, which it knows exist but have not yet been crawled and downloaded."

There were many times I saw examples of this. The top 10 results following a query would sometimes have a basic URL displayed with no title or snippet (or metadata).

Here's an example I used in a presentation from 2004. Look at the bottom result, and you'll see what I mean.

Google Web Images Groups News Froogle more »
french wine Search Advanced Search Preferences

Web Results 1 - 10 of about 7,860,000 for **french wine**. (0.15 seconds)

Tip: Save time by hitting the return key instead of clicking on "search"

News results for **french wine** - [View today's top stories](#)
 French wine worth its weight in gold - [Daily Times](#) - 28 Nov 2004

Wines of France - A guide to French wines
Experience **French Wines** in the USA. [www.onivins.fr](#) [Language] [Legal Note] © Vins de France 2003. © Vins de France 2003.
[www.frenchwinesfood.com/](#) - 8k - [Cached](#) - [Similar pages](#)

French Wine Guide
French Wine Guide is the complete guide to **French wines** with regions of France, Bordeaux region, **wine** and food, **french wine** classification, tasting, winemaking ...
[www.terroir-france.com/](#) - 27k - 28 Nov 2004 - [Cached](#) - [Similar pages](#)

Find luxury French wine tours, France wine tasting vacations, and ...
Find luxury **French wine** tours, private France **wine** tasting vacations, and gourmet travel. ...
Luxury **French Wine** and Culinary Tours. ...
[www.wine-tours-france.com/](#) - 16k - [Cached](#) - [Similar pages](#)

All French wines, Bordeaux, Burgundy wines and Champagne - Buy ...
... **french wines**. On line sales of **French** great vintages, Bordeaux and Burgundy **wines**, red and white **wine**. ... your table. The **French wine** place : ...
[www.french-wine-shop.com/](#) - 17k - [Cached](#) - [Similar pages](#)

[www.french-wine-web.com/](#)
[Similar pages](#)

Sponsored Links

French Wine
Taste & buy the best France has to offer at the Olympia show this Jan.
[www.vivelafrance.co.uk](#)

Wines of France
An original selection of fine wines
Tasting Cartons and good values
[www.winesoffrance.com.fr](#)

French wine UK
Save 10% on **French wine** by the case with guaranteed delivery in the UK
[www.VirginWines.com](#)

French Wine On-line
Vin de Pays d'Oc, Burgundy, Jura, Rhone, Alsace, Bordeaux, Loire etc.
[www.lebonvin.co.uk](#)

French Wine from Tesco
Up to 50% off all Champagne & 25% off all USA and South American **wine**
[www.tesco.com/wine](#)

More than 15 000 wines
Bordeaux, Primeurs, Champagnes
A gift of Champagne with your order

Google Web Images Groups News Froogle more »
 french wine Search Advanced Search Preferences

Web Results 1 - 10 of about 7,860,000 for **french wine**. (0.15 seconds)

Tip: Save time by hitting the return key instead of clicking on "search"

News results for **french wine** - View today's top stories
 **French wine worth its weight in gold** - Daily Times - 28 Nov 2004

Wines of France - A guide to French wines
 Experience **French Wines** in the USA. www.onivins.fr |Language |Legal Note © Vins de France 2003. © Vins de France 2003.
 www.frenchwinesfood.com/ - 8k - [Cached](#) - [Similar pages](#)

French Wine Guide
French Wine Guide is the complete guide to **French wines** with regions of France, Bordeaux region, **wine** and food, **french wine** classification, tasting, winemaking ...
 www.terroir-france.com/ - 27k - 28 Nov 2004 - [Cached](#) - [Similar pages](#)

Find luxury French wine tours, France wine tasting vacations, and ...
 Find luxury **French wine** tours, private France **wine** tasting vacations, and gourmet travel. ...
 Luxury **French Wine** and Culinary Tours. ...
 www.wine-tours-france.com/ - 16k - [Cached](#) - [Similar pages](#)

All French wines, Bordeaux, Burgundy wines and Champagne - Buy ...
 ... **french wines**. On line sales of **French** great vintages, Bordeaux and Burgundy **wines**, red and white **wine**. ... your table. The **French wine** place : ...
 www.french-wine-shop.com/ - 17k - [Cached](#) - [Similar pages](#)

[www.french-wine-web.com/](#)
[Similar pages](#)

Sponsored Links

French Wine
 Taste & buy the best France has to offer at the Olympia show this Jan.
[www.vivelafrance.co.uk](#)

Wines of France
 An original selection of fine wines
 Tasting Cartons and good values
[www.winesoffrance.com.fr](#)

French wine UK
 Save 10% on **French wine** by the case with guaranteed delivery in the UK
[www.VirginWines.com](#)

French Wine On-line
 Vin de Pays d'Oc, Burgundy, Jura, Rhone, Alsace, Bordeaux, Loire etc.
[www.lebonvin.co.uk](#)

French Wine from Tesco
 Up to 50% off all Champagne & 25% off all USA and South American **wine**
[www.tesco.com/wine](#)

More than 15,000 wines
 Bordeaux, Primeurs, Champagnes
 A gift of Champagne with your order

Google is aware of the importance of that page because of the linkage data surrounding it. But no supporting information has been pulled from the page, not even the title tag, as the page obviously hasn't been crawled. (Of course, this can also occur with the evergreen still-happens-all-the-time little blunder when someone leaves the robots.txt file preventing the site from being crawled.)

I highlighted that sentence above in bold for two important reasons:

- Hyperlink analysis can denote the "importance" of a page before it even gets crawled and indexed. Along with bandwidth and politeness, the importance of a page is one of the three primary considerations when plotting the crawl. (We'll dive deeper into hyperlinks and hyperlink-based ranking algorithms in future installments.)
- Every now and again, the "are links still important" debate flares up (and then cools down). Trust me. The answer is yes, links are still important.

I'll just embellish the "politeness" thing a little more as

it's directly connected to the robots.txt file/protocol. All the challenges to crawling the web that I explained 20 years ago still exist today (at a greater scale).

Because crawlers retrieve data at vastly much greater speed and depth than humans, they could (and sometimes do) have a crippling impact on a website's performance. Servers can crash just trying to keep up with the number of rapid-speed requests.

That's why a politeness policy governed on the one hand by the programming of the crawler and the plot of the crawl, and on the other by the robots.txt file is required.

The faster a search engine can crawl new content to be indexed and recrawl existing pages in the index, the fresher the content will be.

Getting the balance right? That's the hard part.

Let's say, purely hypothetically, that Google wanted to keep thorough coverage of news and current affairs and decided to try and crawl the entire New York Times website every day (even every week) without any politeness factor at all. It's most likely that the crawler would use up all their bandwidth. And that would mean that nobody can get to read the paper online because of bandwidth hogging.

Thankfully now, beyond just the politeness factor, we have Google Search Console, where it's possible to manipulate the speed and frequency of which websites are crawled.

What's changed in 32 years of crawling the web?

OK, we've covered a lot of ground as I knew we would.

There have certainly been many changes to both the internet

and the world wide web – but the crawling part still seems to be impeded by the same old issues.

That said, a while back, I saw a presentation by Andrey Kolobov, a researcher in the field of machine learning at Bing. He created an algorithm to do a balancing act with the bandwidth, politeness and importance issue when plotting the crawl.

I found it highly informative, surprisingly straightforward and pretty easily explained. Even if you don't understand the math, no worries, you'll still get an indication of how he tackles the problem. And you'll also hear the word "importance" in the mix again.

Basically, as I explained earlier about URLs on the frontier of the crawl, hyperlink analysis is important before you get crawled, indeed may well be the reason behind how quickly you get crawled. You can watch the short video of his presentation [here](#).

Now let's wind up with what's occurring with the internet right now and how the web, internet, 5G and enhanced content formats are cranking up.

Structured data

The web has been a sea of unstructured data from the get-go. That's the way it was invented. And as it still grows exponentially every day, the challenge the search engines have is having to crawl and recrawl existing documents in the index to analyze and update if any changes have been made to keep the index fresh.

It's a mammoth task.

It would be so much easier if the data were structured. And so much of it actually is, as structured databases drive so many websites. But the content and the presentation are separated,

of course, because the content has to be published purely in HTML.

There have been many attempts that I've been aware of over the years, where custom extractors have been built to attempt to convert HTML into structured data. But mostly, these attempts were very fragile operations, quite laborious and totally error-prone.

Something else that has changed the game completely is that websites in the early days were hand-coded and designed for the clunky old desktop machines. But now, the number of varying form factors used to retrieve web pages has hugely changed the presentation formats that websites must target.

As I said, because of the inherent challenges with the web, search engines such as Google are never likely ever to be able to crawl and index the entire world wide web.

So, what would be an alternative way to vastly improve the process? What if we let the crawler continue to do its regular job and make a structured data feed available simultaneously?

Over the past decade, the importance and usefulness of this idea have grown and grown. To many, it's still quite a new idea. But, again, Pinkerton, WebCrawler inventor, was way ahead on this subject 20 years ago.

He and I discussed the idea of domain-specific XML feeds to standardize the syntax. At that time, XML was new and considered to be the future of browser-based HTML.

It's called extensible because it's not a fixed format like HTML. XML is a "metalanguage" (a language for describing other languages which lets you design your own customized markup languages for limitless diverse types of documents). Various other approaches were vaunted as the future of HTML but couldn't meet the required interoperability.

However, one approach that did get a lot of attention is known as MCF (Meta Content Framework), which introduced ideas from the field of knowledge representation (frames and semantic nets). The idea was to create a common data model in the form of a directed labeled graph.

Yes, the idea became better known as the semantic web. And what I just described is the early vision of the knowledge graph. That idea dates to 1997, by the way.

All that said, it was 2011 when everything started to come together, with schema.org being founded by Bing, Google, Yahoo and Yandex. The idea was to present webmasters with a single vocabulary. Different search engines might use the markup differently, but webmasters had to do the work only once and would reap the benefits across multiple consumers of the markup.

OK – I don't want to stray too far into the huge importance of structured data for the future of SEO. That must be an article of its own. So, I'll come back to it another time in detail.

But you can probably see that if Google and other search engines can't crawl the entire web, the importance of feeding structured data to help them rapidly update pages without having to recrawl them repeatedly makes an enormous difference.

Having said that, and this is particularly important, you still need to get your unstructured data recognized for its E-A-T (expertise, authoritativeness, trustworthiness) factors before the structured data really kicks in.

Cloud computing

As I've already touched on, over the past four decades, the internet has evolved from a peer-to-peer network to overlaying the world wide web to a mobile internet revolution, Cloud computing, the Internet of Things, Edge Computing, and 5G.

The shift toward Cloud computing gave us the industry phrase “the Cloudification of the internet.”

Huge warehouse-sized data centers provide services to manage computing, storage, networking, data management and control. That often means that Cloud data centers are located near hydroelectric plants, for instance, to provide the huge amount of power they need.

Edge computing

Now, the “Edgeification of the internet” turns it all back around from being further away from the user source to being right next to it.

Edge computing is about physical hardware devices located in remote locations at the edge of the network with enough memory, processing power, and computing resources to collect data, process that data, and execute it in almost real-time with limited help from other parts of the network.

By placing computing services closer to these locations, users benefit from faster, more reliable services with better user experiences and companies benefit by being better able to support latency-sensitive applications, identify trends and offer vastly superior products and services. IoT devices and Edge devices are often used interchangeably.

5G

With 5G and the power of IoT and Edge computing, the way content is created and distributed will also change dramatically.

Already we see elements of virtual reality (VR) and augmented reality (AR) in all kinds of different apps. And in search, it will be no different.

AR imagery is a natural initiative for Google, and they’ve

been messing around with 3D images for a couple of years now just testing, testing, testing as they do. But already, they're incorporating this low-latency access to the knowledge graph and bringing in content in more visually compelling ways.

During the height of the pandemic, the now "digitally accelerated" end-user got accustomed to engaging with the 3D images Google was sprinkling into the mix of results. At first it was animals (dogs, bears, sharks) and then cars.

Last year Google announced that during that period the 3D featured results interacted with more than 200 million times. That means the bar has been set, and we all need to start thinking about creating these richer content experiences because the end-user (perhaps your next customer) is already expecting this enhanced type of content.

If you haven't experienced it yourself yet (and not everyone even in our industry has), here's a very cool treat. In this video from last year, Google introduces famous athletes into the AR mix. And superstar athlete Simone Biles gets to interact with her AR self in the search results.

IoT

Having established the various phases/developments of the internet, it's not hard to tell that everything being connected in one way or another will be the driving force of the future.

Because of the advanced hype that much technology receives, it's easy to dismiss it with thoughts such as IoT is just about smart lightbulbs and wearables are just about fitness trackers and watches. But the world around you is being incrementally reshaped in ways you can hardly imagine. It's not science fiction.

IoT and wearables are two of the fastest-growing technologies

and hottest research topics that will hugely expand consumer electronics applications (communications especially).

The future is not late in arriving this time. It's already here.

We live in a connected world where billions of computers, tablets, smartphones, wearable devices, gaming consoles and even medical devices, indeed entire buildings are digitally processing and delivering information.

Here's an interesting little factoid for you: it's estimated that the number of devices and items connected to IoT already eclipses the number of people on earth.

Back to the SEO future

We'll stop here. But much more to come.

I plan to break down what we now know as search engine optimization in a series of monthly articles scoping the foundational aspects. Although, the term "SEO" wouldn't enter the lexicon for some while as the cottage industry of "doing stuff to get found at search engine portals" began to emerge in the mid-to-late 1990s.

Until then – be well, be productive and absorb everything around you in these exciting technological times. I'll be back again with more in a few weeks.

Opinions expressed in this article are those of the guest author and not necessarily Search Engine Land. Staff authors are listed here.

New on Search Engine Land

About The Author





Mike Grehan is an SEO pioneer (online since 1995), author, world-traveler and keynote speaker, Champagne connoisseur and consummate drinking partner to the global digital marketing community. Former publisher of SearchEngineWatch.com and ClickZ.com, and producer of the industry's largest search and social marketing event, SES Conference & Expo. Proud to have

been chairman of SEMPO the largest global trade association for search marketers. And equally proud to be SVP of corporate communications, NP Digital.

<https://searchengineland.com/web-crawling-history-383908>