search.gov

How Search Engines Index Your Websites

May 29, 2018 | Dawn Pointer McCleskey

Overview

How Search Engines Work

XML Sitemaps

Robots.txt Files

What to do about multiple publishing platforms

Relationship between indexing and search

What is a search engine?

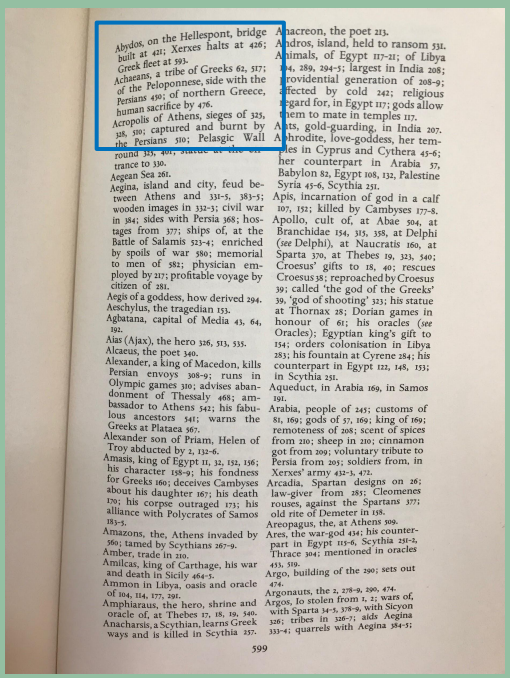
I’d like to start out with some definitions so that we’re all on a the same page.

Index:

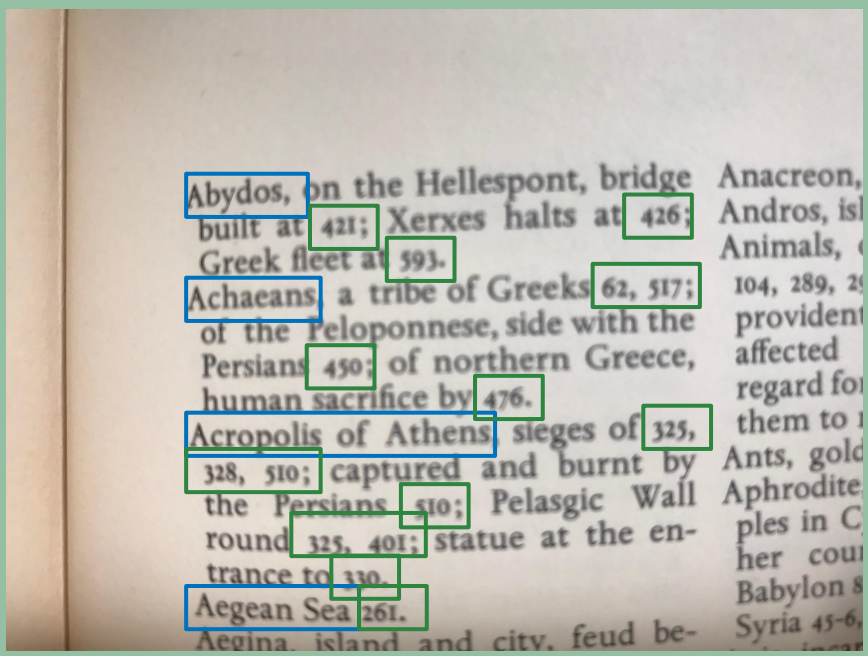
1a. a device (such as the pointer on a scale or the gnomon of a sundial) that serves to indicate a value or quantity

2a. a list of items (such as topics or names) treated in a printed work that gives for each item the page number where it may be found

- https://www.merriam-webster.com/dictionary/index



This is a page from the index of our copy of Herodotus, and as you know it shows on what pages we'll find certain information. Let's zoom in on the top portion here.



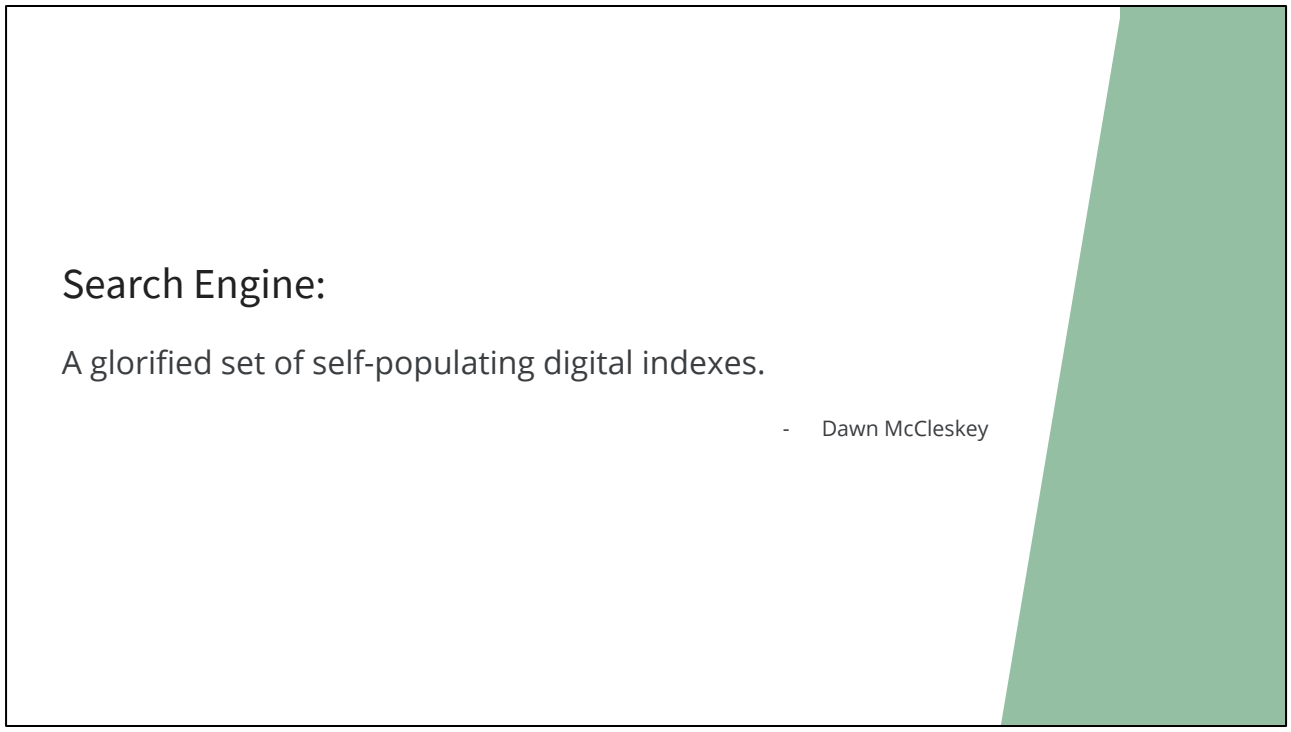
Some indexes show the key word and then just page numbers, but you'll see this shows not only the keyword but also some context that allows you to make an informed decision about which page you want to turn to. This is old school search results with snippets. Each of these entries shows metadata about the page. [Next slide]

Metadata:

data that provides information about other data

- https://www.merriam-webster.com/dictionary/metadata

Metadata, of course is data about other data. In this case we had information on a page, and we had that information's topic, subtopic, and the number of the page. None of this is the information on the page, if we want that we'll have to go read it directly.



This is my very simplified definition of a search engine - the thing that makes them different from a regular database is that they self populate, and so they appear very magical. But other than the scale, sophistication of results presentation, and how information gets into them, using Google isn't really different than doing a Find In All Sheets search in Excel. Just like the Wizard of Oz, there are people behind the curtain using technology to make it look like magic. So let's draw back that curtain.

Yes but how does a search engine *work*?

1. Discover URLS
2. Parse data to insert into the index
3. Support search of the index

There are two sides to any search engine, indexing content, which consists of discovering what content is available, and building a useful index of that content. And then they also allow queries against the index so that end users can discover the content for themselves.

Crawling

XML Sitemaps

Feeds

There are two primary ways that search engines discover your content. Crawling and Sitemaps.

Crawling is when the search bot opens a page, detects all the links on that page, makes a note of them, and then goes to those pages as well.

XML Sitemaps are machine-friendly lists of the items on a website, and the bot reads the lists and notes all the URLs. https://search.gov/blog/sitemaps.html

Both of these result in the search engine knowing what exists on your website. One very important thing to note is that Google and Bing have clearly stated that they don't crawl everything, or the entirety of any given site, and that they don't index everything that they do crawl through. References:
https://support.google.com/webmasters/answer/34441?hl=en,
https://www.bing.com/webmaster/help/why-is-my-site-not-in-the-index-2141dfab

Google also can discover content via Feeds - yes! Feeds! I won't cover this method today, but there will be a link in the slides if you want to learn more.
https://support.google.com/webmasters/answer/178852

Control access to your content with ...

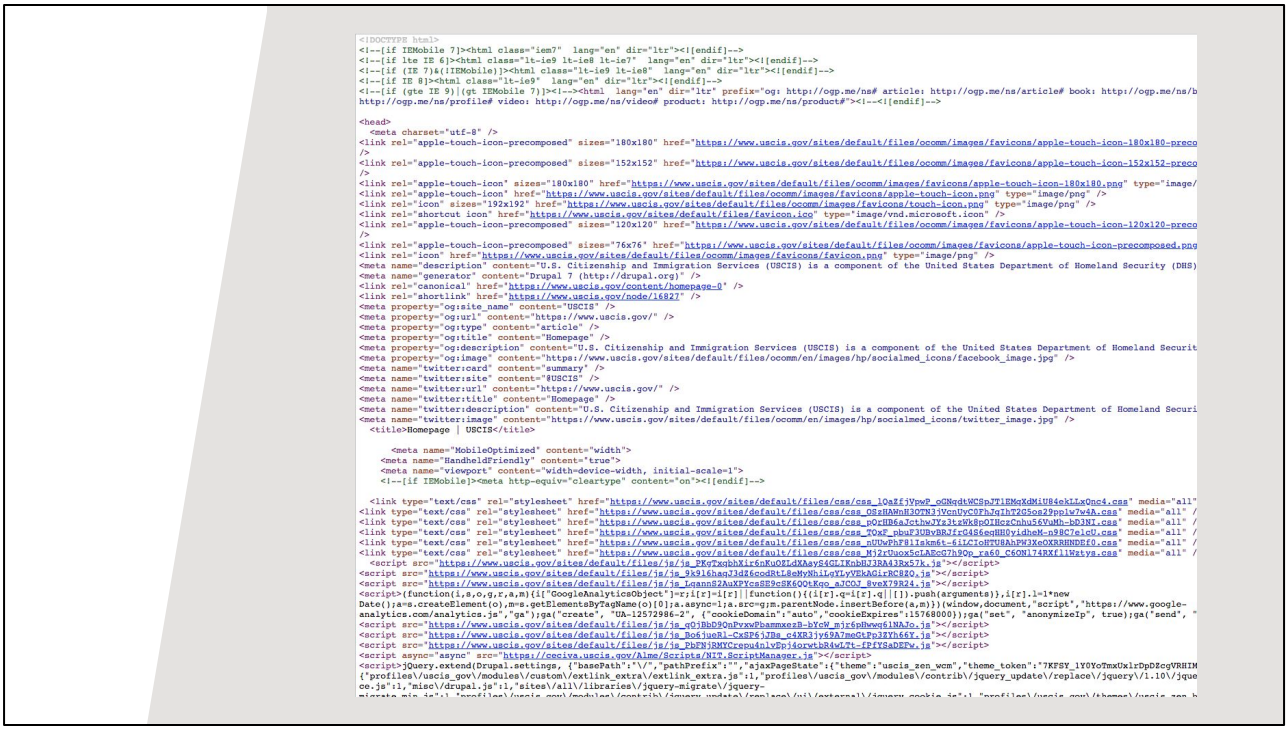
robots.txt

meta robots tags

You can indicate whether you want portions of your site indexed or not using Robots settings, and we'll talk about these in detail later.
To mark entire folders as not to be indexed, use a robots.txt file
For individual pages, you can list those in the robots.txt file if you want, or you can use a robots meta tag in the head of your page.

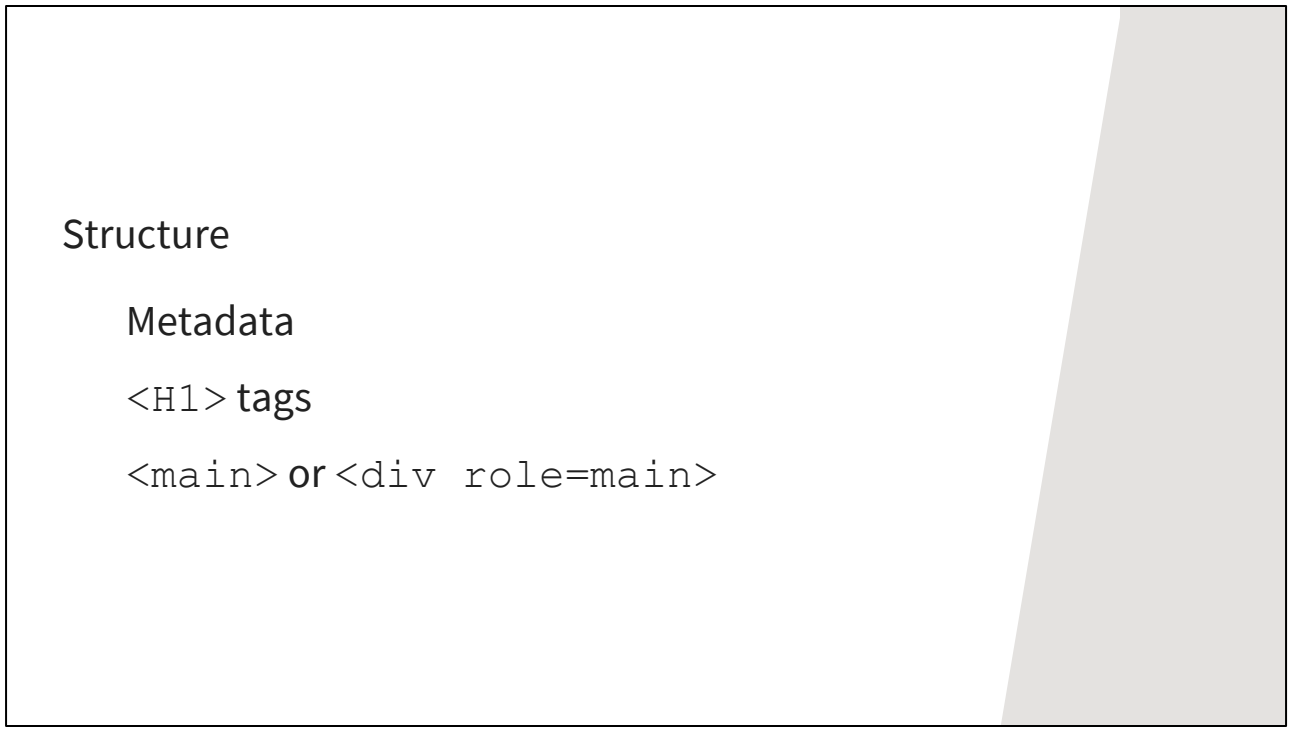
More on both of these later.

So, the bot has gone through your site and collected your URLs



Now it's going to parse the data on each of those pages. The bot requests the page, reads the code, and sorts through what's there to see what can be used, according to the developers who programmed the bot. What links are on the page? What text based content is there? Images or other content types? They look to see what's there, sort it out, map it into their own model, and save the information to their indexes. That's the actual indexing stage, when the parsed data gets saved.

As some of you have heard me say before, computers are stupid and can only do what they've been told to do. So far, at least. Parsing of content is helped by...



Structure.

The more explicit structure you can put into your pages, the easier it will be for the computer to interpret your intent and present good results to searchers.
?? Do you have metadata in tags in your page heads, or inline in the tags within the body? I won't talk much about structured markup today, but if you have questions about that, let's follow up afterwards. If you use a content management system, there are plugins that will insert structured markup for you.
?? Do you use header tags, the H tags, in your page in the proper way, with the page title in H1, H2 containing section headings, etc.? I've seen page titles in H2s, while the site section name is in the H1, and there are plenty of pages still in government that have the title and sections just as bold text in the body. You'll get better results in search if your title is in the H1.
?? Are you using a "main" element in your body tag? The body tag of course includes all the visible stuff on the page, but we don't want headers, footers, and navigation elements indexed along with the main content of a given page. So we wrap that portion of the html body in a main tag, letting bots know that anything outside of this is fluff and can be ignored.
→ watch out that you've implemented this properly. The other day I was trying to figure out why a particular search result was so hidden for a page we had indexed, and it was because the only thing inside the main element was the H1 tag. The bulk of the important content for that page was getting ignored.

High quality content

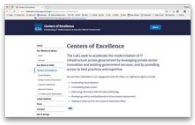
Unique to page

Plain language, well written

Useful

The higher quality the content, the better your SEO ranking will be, because the search engines will be better able to match searcher needs with your content. What does High Quality Content mean?

- It should be unique to the page. This includes the main content of the page, but also title tags and meta descriptions in the head
- Write in plain language, using the words your audiences use in addition to whatever technical terms or acronyms you may need to use. You'll want to go easy on the jargon, though. This is win-win because it'll be easier for the search engine to match your content with searcher needs, and once they get to the page, they'll be able to understand it!
- It should be well written, too. In the age of Grammarly and other quality checkers, it's no surprise that search engines would start running these tools over the content in their indexes and scoring pages accordingly. I know some agencies provide plain language and quality review tools to their staff, so you should ask around. This applies to new pages as well as ones that have been around a while.
- And finally, is your content useful? Search engines track what people are looking for, and what they click on from results, and analyzing what makes a particular item click-worthy. If a page appears redundant, too fluffy, or incomprehensible, it shouldn't be too surprising if it doesn't end up getting indexed.



Government Website

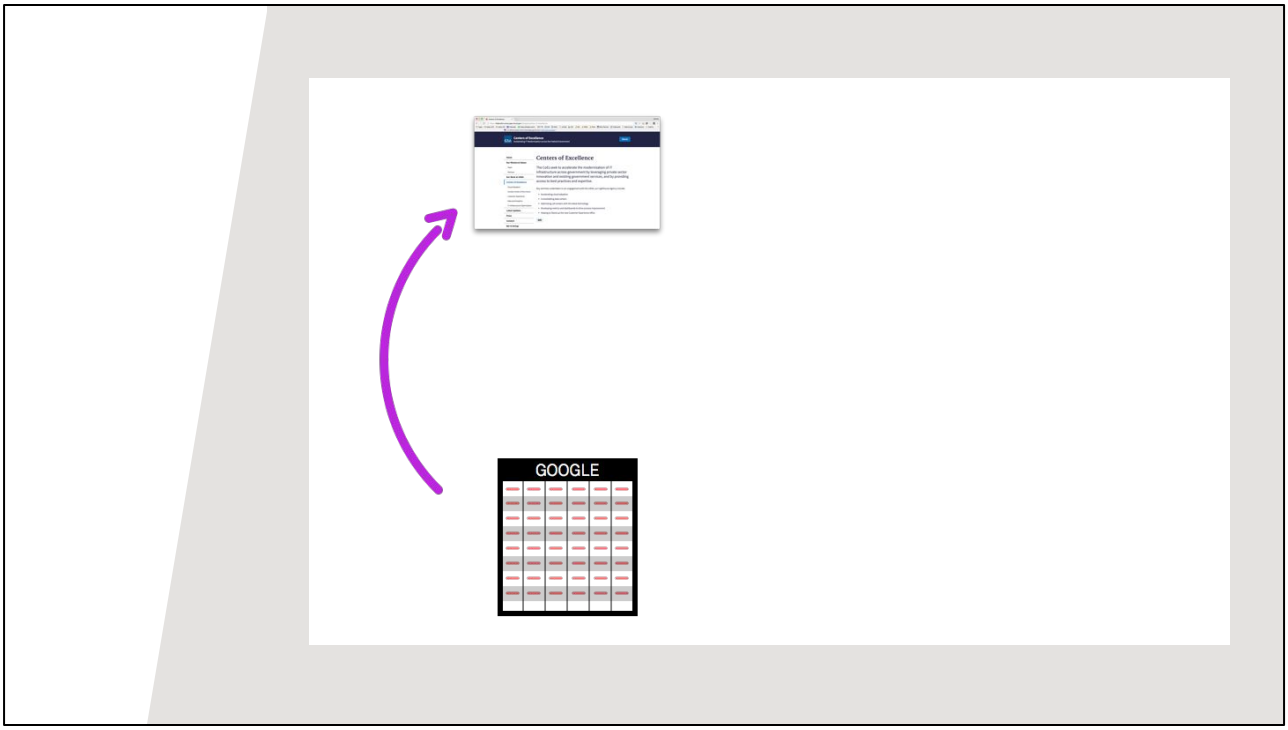
So let’s recap what we’ve gone over so far, and I’ll take questions in a minute.

This is your website

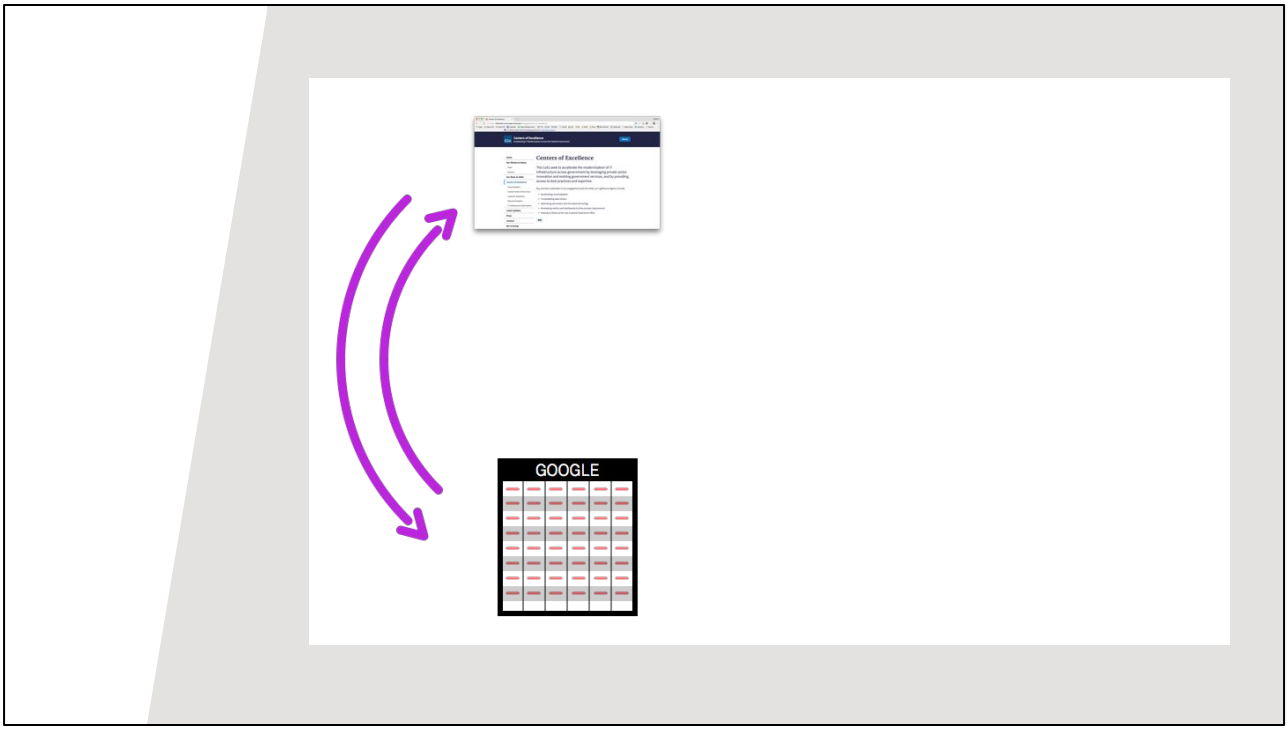
Government Website

Search Engine

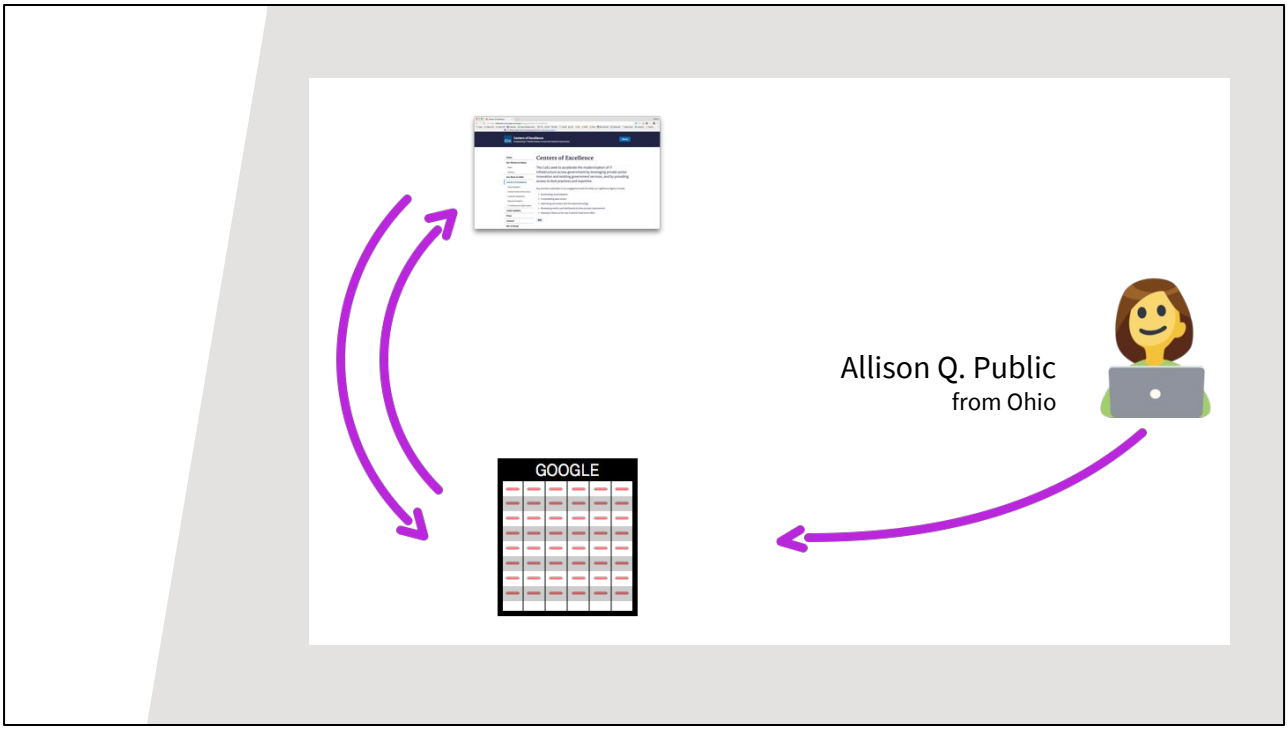
And here is Google over in some other part of the web.



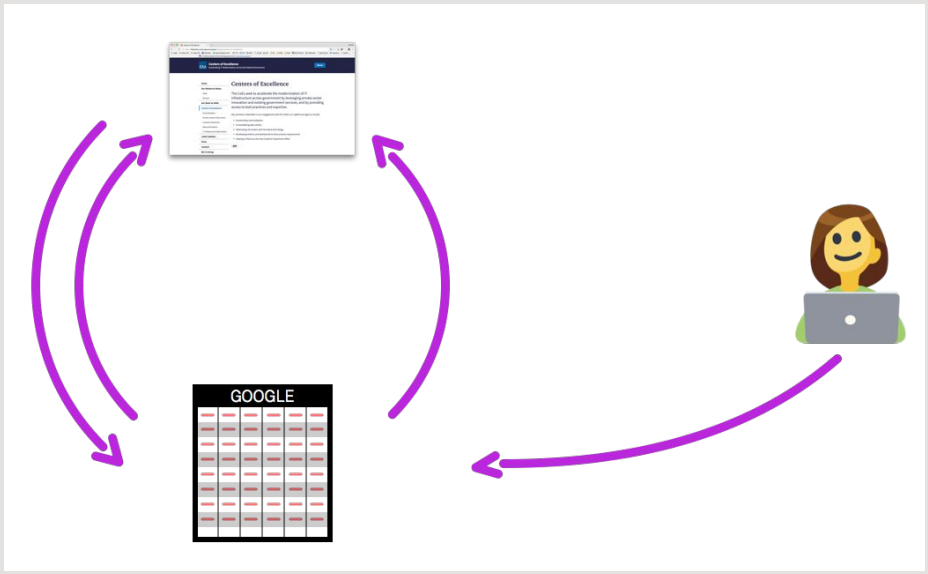
Googlebot visits your website to review your sitemap and crawl around



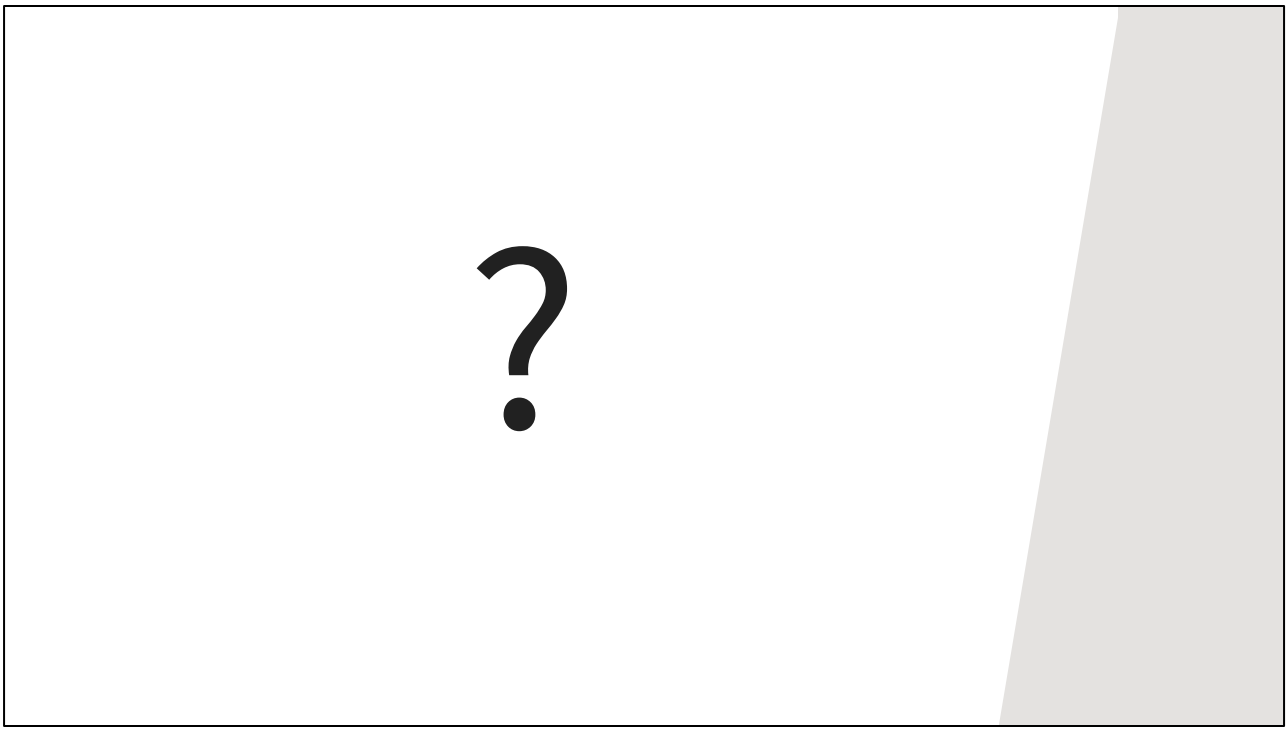
They parse your data and decide what of it they'll bring into their indexes.



This is Allison, she wants to get a passport. So she goes to google and searches for passport application.



She sees a bunch of results, clicks on one from travel.state.gov and is brought over to that site to view the information.



Do we have any questions so far?

Crawling

Make your site more crawler friendly by...

Avoiding crawler traps

Using `rel=canonical` link tags

```
<head>
<link rel="canonical" href="https://agency.gov/topic1" />
</head>
```

JavaScript sites: avoid URLs with /#/, use /#!/

We're not going to talk in depth about crawling today, but you do want to make your site as crawlable as possible:

Crawler traps occur when your site can generate an infinite number of URLs, and this usually happens when any given URL can have parameters appended to it, like tags, referring pages, google tag manager tokens, and so on. Each one of those URLs will look like a different URL, and the crawler will open it, note the links on it, which then themselves, once followed, will have these additional parameters on them, and the crawler won't be able to figure out what constitutes the entirety of a site. Most, if not all, search engines will stop working if they detect a crawler trap. Reference: https://support.archive-it.org/hc/en-us/articles/208332943-Identify-and-avoid-crawler-traps-

You can help avoid crawler traps as well as alleviate any other potential confusion over which version of a URL is the version of record by inserting a canonical link in the head of the page. This particular example tells search engines that even though this page can be accessed via www.agency.gov/topic1/, this version without the subdomain and the trailing slash should be the one indexed. This helps with crawler traps, too, because it says "disregard all those parameters and index just the real URL." Reference: https://yoast.com/rel-canonical/

And if your site is built with JavaScript, note that even though Search Engines can now render your pages to see the content, they're still ignoring anything past a hash sign in the URL, because within-page anchor links use hashes as well. So you'll want either to get rid of the hash sign entirely, or convert it to a hash-bang, where there's an exclamation point following. Googlebot does now handle hash-bang URLs. Reference: https://www.hobo-web.co.uk/javascript-seo/

XML Sitemaps

So let's talk in more detail about XML sitemaps.

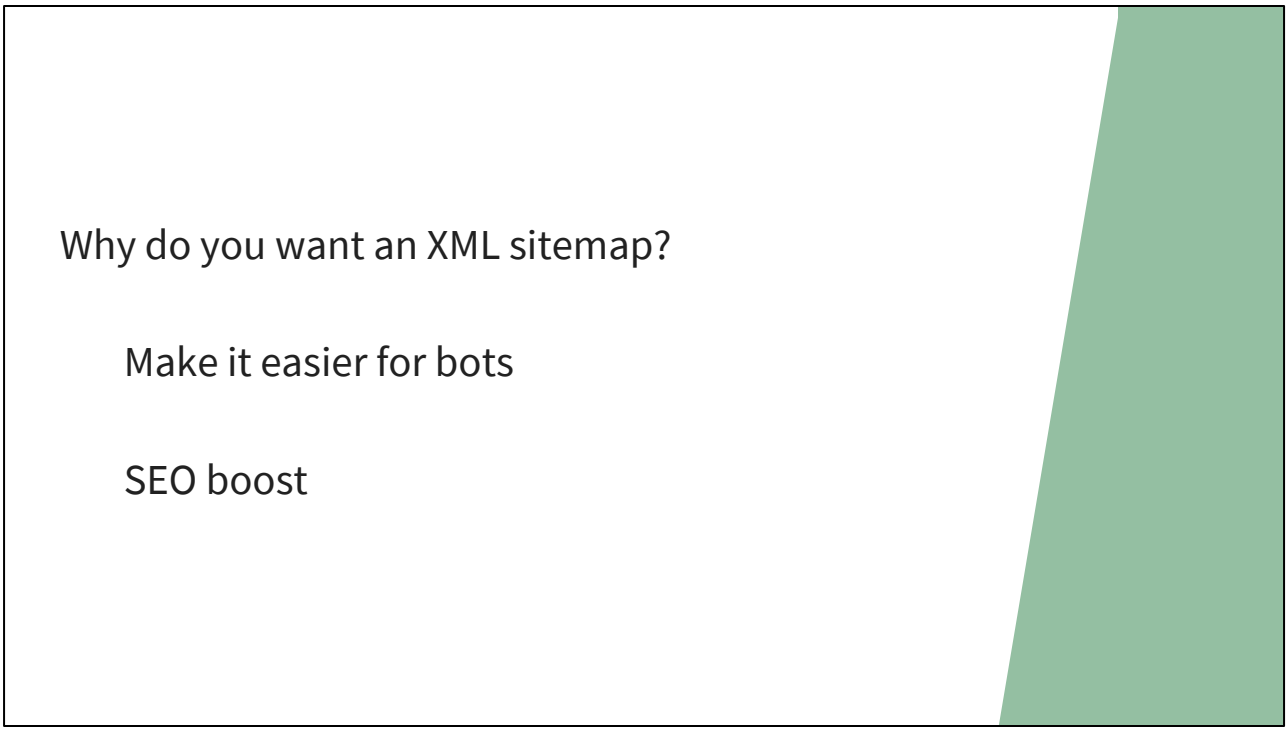
A sitemap is a list of items on a website, hopefully stating when each was last modified, and possibly indicating their relative import.

It is NOT a shopping list for Google of all the things they should pick up from your site.

It's more like a weekly specials flyer saying Here's what we have!

There's a philosophical debate here. Do you include only the most important pages?, or all your stuff?

- Googlebot will follow its programming to determine whether to index your content
- Search.gov will take all your stuff
    - We don't editorialize for you
    - We have a different mission, which is to make it easier to find things on a particular website, not across the whole web
    - And because of this, we have a much smaller universe to cover than Google, so we don't have to be as conservative with our index size.



You want an XML sitemap to make it easier for bots to discover your content. It will cut the URL discovery time down significantly.

As a reward for helping them, and for showing you can present your content in an organized way, Google gives a bit of a ranking boost

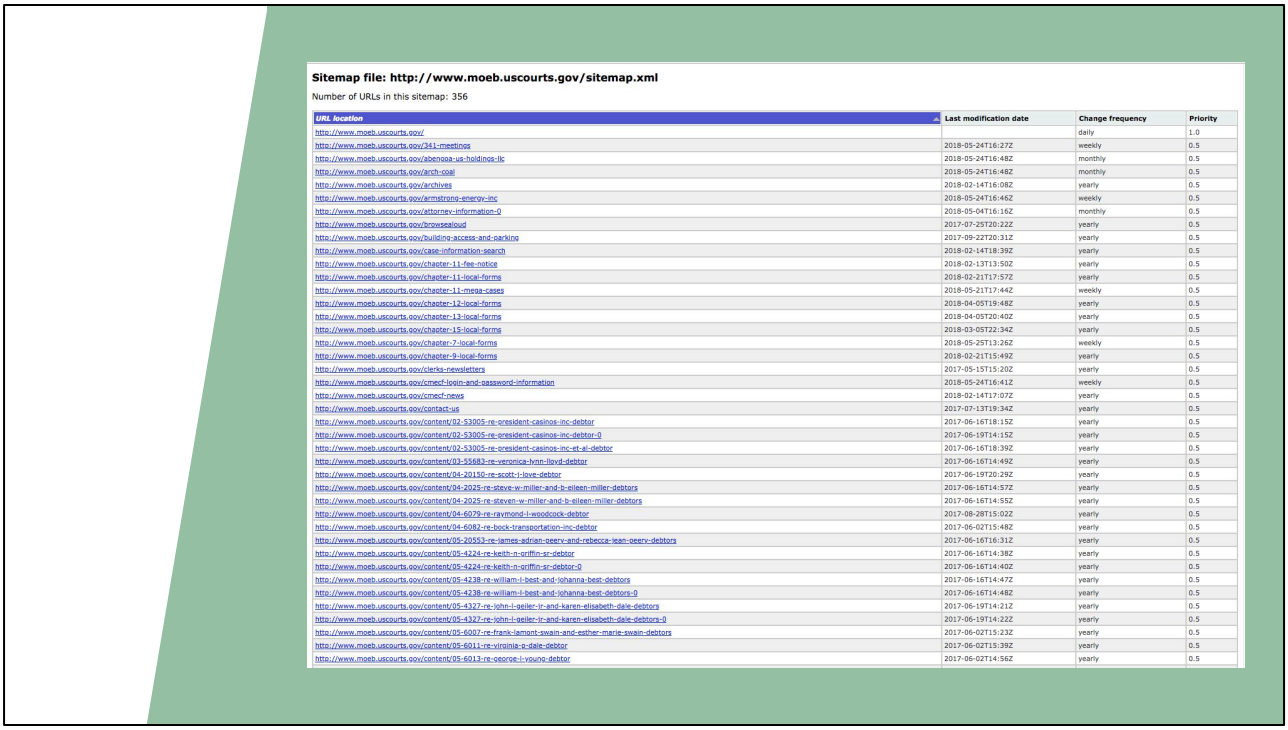
XML Sitemap Protocol

XML format file
Listed in the robots.txt file
For the location where the sitemap file is
Listing clean URLs
Date each file last modified
Optional: file change frequency & priority

https://www.sitemaps.org/protocol.html

And that's it, it's pretty basic, but each area is tricky.

Your sitemap might look like this - this is our search.gov sitemap, generated by a jekyll tool in github pages, and it includes just our URLs and last modified date, or...



Sitemap file: http://www.moeb.uscourts.gov/sitemap.xml

Number of URLs in this sitemap: 356

URL location	Last modification date	Change frequency	Priority
http://www.moeb.uscourts.gov/		daily	1.0
http://www.moeb.uscourts.gov/341-meetings	2018-05-24T16:27Z	weekly	0.5
http://www.moeb.uscourts.gov/abengoa-us-holdings-llc	2018-05-24T16:48Z	monthly	0.5
http://www.moeb.uscourts.gov/arch-coal	2018-05-24T16:48Z	monthly	0.5
http://www.moeb.uscourts.gov/archives	2018-02-14T16:08Z	yearly	0.5
http://www.moeb.uscourts.gov/armstrong-energy-inc	2018-05-24T16:46Z	weekly	0.5
http://www.moeb.uscourts.gov/attorney-information-0	2018-05-04T16:16Z	monthly	0.5
http://www.moeb.uscourts.gov/browsealoud	2017-07-25T20:22Z	yearly	0.5
http://www.moeb.uscourts.gov/building-access-and-parking	2017-09-22T20:31Z	yearly	0.5
http://www.moeb.uscourts.gov/case-information-search	2018-02-14T18:39Z	yearly	0.5
http://www.moeb.uscourts.gov/chapter-11-fee-notice	2018-02-13T13:50Z	yearly	0.5
http://www.moeb.uscourts.gov/chapter-11-local-forms	2018-02-21T17:57Z	yearly	0.5
http://www.moeb.uscourts.gov/chapter-11-mega-cases	2018-05-21T17:44Z	weekly	0.5
http://www.moeb.uscourts.gov/chapter-12-local-forms	2018-04-05T19:48Z	yearly	0.5
http://www.moeb.uscourts.gov/chapter-13-local-forms	2018-04-05T20:40Z	yearly	0.5
http://www.moeb.uscourts.gov/chapter-15-local-forms	2018-03-05T22:34Z	yearly	0.5
http://www.moeb.uscourts.gov/chapter-7-local-forms	2018-05-25T13:26Z	weekly	0.5
http://www.moeb.uscourts.gov/chapter-9-local-forms	2018-02-21T15:49Z	yearly	0.5
http://www.moeb.uscourts.gov/clerks-newsletters	2017-05-15T15:20Z	yearly	0.5
http://www.moeb.uscourts.gov/cmecf-login-and-password-information	2018-05-24T16:41Z	weekly	0.5
http://www.moeb.uscourts.gov/cmecf-news	2018-02-14T17:07Z	yearly	0.5
http://www.moeb.uscourts.gov/contact-us	2017-07-13T19:34Z	yearly	0.5
http://www.moeb.uscourts.gov/content/02-53005-re-president-casinos-inc-debtor	2017-06-16T18:15Z	yearly	0.5
http://www.moeb.uscourts.gov/content/02-53005-re-president-casinos-inc-debtor-0	2017-06-19T14:15Z	yearly	0.5
http://www.moeb.uscourts.gov/content/02-53005-re-president-casinos-inc-et-al-debtor	2017-06-16T18:39Z	yearly	0.5
http://www.moeb.uscourts.gov/content/03-55683-re-veronica-lynn-lloyd-debtor	2017-06-16T14:49Z	yearly	0.5
http://www.moeb.uscourts.gov/content/04-20150-re-scott-i-love-debtor	2017-06-19T20:29Z	yearly	0.5
http://www.moeb.uscourts.gov/content/04-2025-re-steve-w-miller-and-b-eileen-miller-debtors	2017-06-16T14:57Z	yearly	0.5
http://www.moeb.uscourts.gov/content/04-2025-re-steven-w-miller-and-b-eileen-miller-debtors	2017-06-16T14:55Z	yearly	0.5
http://www.moeb.uscourts.gov/content/04-6079-re-raymond-l-woodcock-debtor	2017-08-28T15:02Z	yearly	0.5
http://www.moeb.uscourts.gov/content/04-6082-re-bock-transportation-inc-debtor	2017-06-02T15:48Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-20553-re-james-adrian-peery-and-rebecca-lean-peery-debtors	2017-06-16T16:31Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-4224-re-keith-n-griffin-sr-debtor	2017-06-16T14:38Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-4224-re-keith-n-griffin-sr-debtor-0	2017-06-16T14:40Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-4238-re-william-l-best-and-johanna-best-debtors	2017-06-16T14:47Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-4238-re-william-l-best-and-johanna-best-debtors-0	2017-06-16T14:48Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-4327-re-john-l-geiler-jr-and-karen-elisabeth-dale-debtors	2017-06-19T14:21Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-4327-re-john-l-geiler-jr-and-karen-elisabeth-dale-debtors-0	2017-06-19T14:22Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-6007-re-frank-lamont-swain-and-esther-marie-swain-debtors	2017-06-02T15:23Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-6011-re-virginia-p-dale-debtor	2017-06-02T15:39Z	yearly	0.5
http://www.moeb.uscourts.gov/content/05-6013-re-george-i-young-debtor	2017-06-02T14:56Z	yearly	0.5

… it might look like this - this was generated by the Drupal xml sitemaps module - you can see the change frequency and priority settings here

Now let’s look at the elements of the protocol, and how they can go wrong.

XML format file

```
<?xml version="1.0" encoding="UTF-8"?>
</xml>
```

First it needs to be an XML format file. Yes, there are some alternative formats mentioned in the protocol, but XML is the standard.

So what could go wrong here?

First, it needs an opening XML declaration, with the version stated.
Second, you need to close the XML tag

Again a tool should handle this for you, but I've seen sitemaps that were tool generated but were still missing these features. If you don't do these things, the poor stupid search engine bots won't know what they're looking at.

Listed in the `robots.txt` file

```
Sitemap: https://search.gov/sitemap.xml
```

It should be listed in your robots.txt file, and we'll talk in detail about robots.txt in the next section.

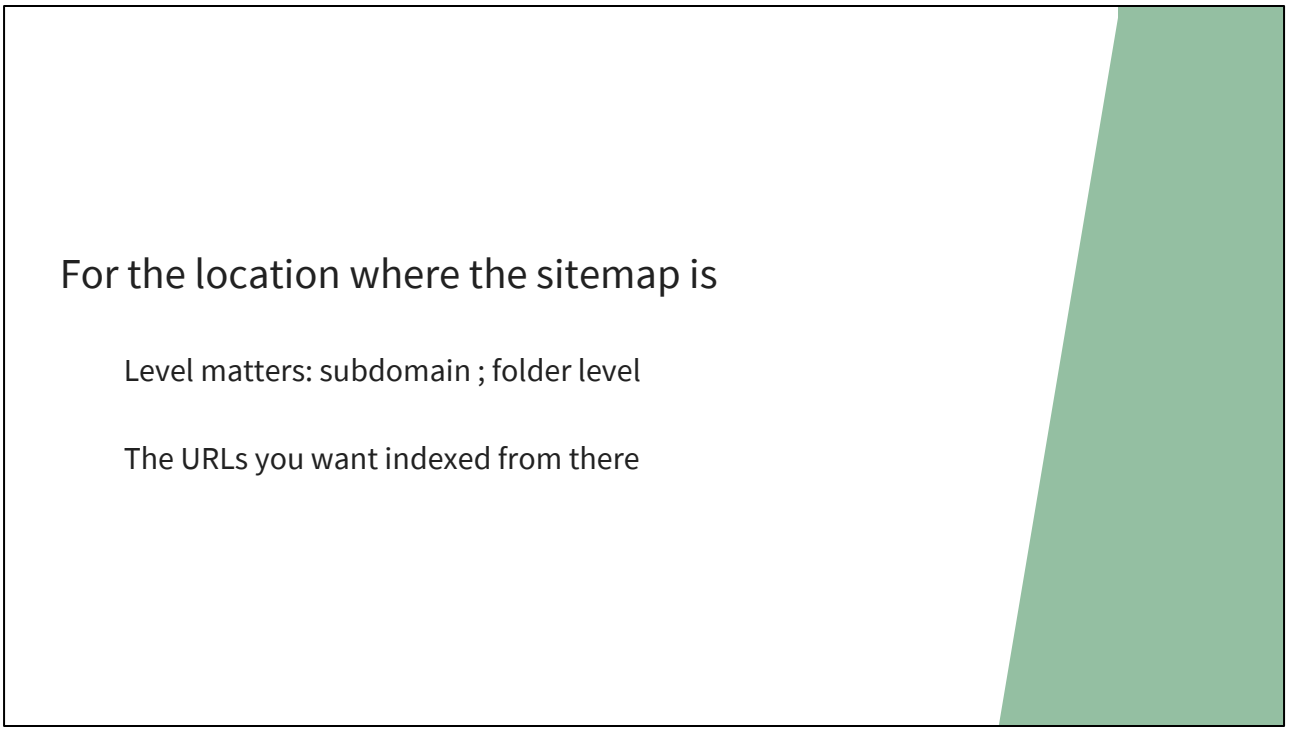
Search engines are going to check your robots.txt file first, and the line for the Sitemap entry should look exactly like this. Sitemap colon space the full path.

In theory this means that you could put the sitemap anywhere, but search engines would still find it because it's listed here. That's true, but violates the location element of the protocol, which we'll look at next.

Don't list the sitemap file as an Allow line.

Do list multiple sitemaps if you have multiple publishing platforms for the subdomain, one on each line.

Don't list multiple sitemaps for multiple domains or subdomains on the same robots.txt. More on that later.

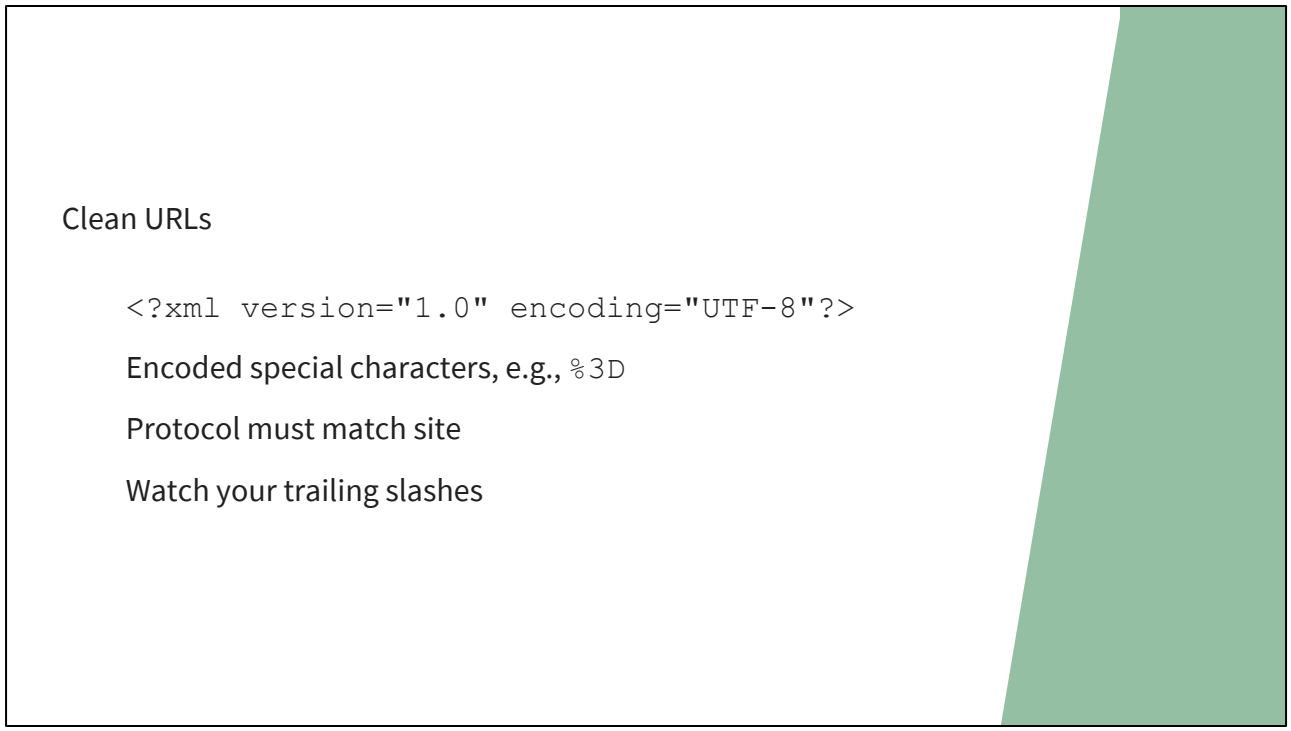


The sitemap should include files for the location where the sitemap is. If a sitemap is at www.agency.gov/sitemap.xml, then it's only supposed to have URLs for the www subdomain. Blog.agency.gov should have its own sitemap at blog.agency.gov/sitemap.xml.

Similarly, if you put your sitemap in a folder, like www.agency.gov/sitemaps/sitemap.xml, according to the protocol it would only have URLs within www.agency.gov/sitemaps/

You also want to include the URLs you want indexed from that location.

- For us, Search.gov, please list all your URLs you want us to index, at least for seeding the index.
- For commercial engines, it's your call, especially if you're happy with your crawl coverage
- For sure, don't include URLs from folders that you've excluded from indexing on your robots.txt file.



Clean URLs

If you're using a content management system, and a plugin to generate your sitemap, this should be taken care of for you.

First, the file needs to be UTF-8 encoded,
and you need to declare this in your opening XML tag.

Any special characters in the URLs need to be encoded

The protocol of the urls in the sitemap need to match the protocol of the site itself. This comes up a lot if the sitemap was created before the site moved to https. So the sitemap file is https, the pages on the site are https, but the pages as listed on the sitemap are still http. This makes double work for the search engine, because even though your 301 redirect might be getting a user to the right place, the search engine will create a record for the http version and then have to update it right away to the https version. Same goes for any URL on your sitemap that redirects to another URL.

Finally, if you have a content management system, it's likely you have directory landing pages that are accessible via agency.gov/section/index.html, agency.gov/section/ and agency.gov/section - each of these will be seen by bots as different urls, so this is a time to decide which version is the preferred version, list that as the canonical URL in the head of the page, and also set that version to be the one included in the sitemap.

Optional fields

- `<Lastmod>` - really helpful!
- `<ChangeFreq>`
- `<Priority>`

If your system stores a date for when pages were last edited, include it on the sitemap in a last mod field. It will indicate to the bot if there's any work to be done on that item since the last time it came to the site.
It also can be used in search and filtering of results, if you don't have publication dates in your page metadata.

If you don't include it, it will be harder for the search engine to pick up updates to the pages, and then we get desperate calls about a leadership bio page that's still showing the name and info about the person who had a very public departure the previous week.

Change frequency isn't really used anymore, mostly because people are people and tried to game the system by saying that page updated daily but there wouldn't be any actual changes to the page for months and months. This isn't a good use of the bots' resources, so the focus is on the lastmod date.

Priority is marginally useful. Because Google and Bing are comparing content across many many sites when they're compiling a list of search results for a given query, knowing that one page on a site is more important than another page on that same site isn't all that helpful. For them, high quality content will be more useful as you compete with other sites for ranking. For us, we're thinking of using this field, because when we're compiling search results, we're usually looking just within one domain, so knowing the relative importance of pages would be informative.

Maximum sitemap size: 50,000 URLs or 50MB

For larger sites, use a sitemap index

- Location specific
- Clean URLs

Google and Bing got together and decided that there should be a maximum size for a given sitemap. In the last year or two they agreed to raise the maximum to 50,000 URLs or 50MB. Anything bigger than this is going to need multiple sitemaps, which will be listed on a sitemap index file.

Most government websites are very large, so this probably applies to you.

The sitemap index should meet all the same standards as an individual sitemap, particularly in that it should be location specific and the urls on it should be clean.

General resources on sitemaps from Google:
https://support.google.com/webmasters/answer/156184
https://support.google.com/webmasters/answer/183668

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
 <sitemap>
  <loc>http://www.agency.gov/sitemap1.xml.gz</loc>
  <lastmod>2017-10-01T18:23:17+00:00</lastmod>
 </sitemap>
 <sitemap>
  <loc>http://www.agency.gov/sitemap2.xml.gz</loc>
  <lastmod>2018-01-01</lastmod>
 </sitemap>
</sitemapindex>
```

Here's an example, you can see that we've got the opening xml declaration, and then after that, instead of a URLset tag, we have a sitemapindex tag. And instead of a url tag with location and lastmod metadata, we have a sitemap tag with the location and lastmod.

Sitemapindex example modified from https://www.sitemaps.org/protocol.html#index

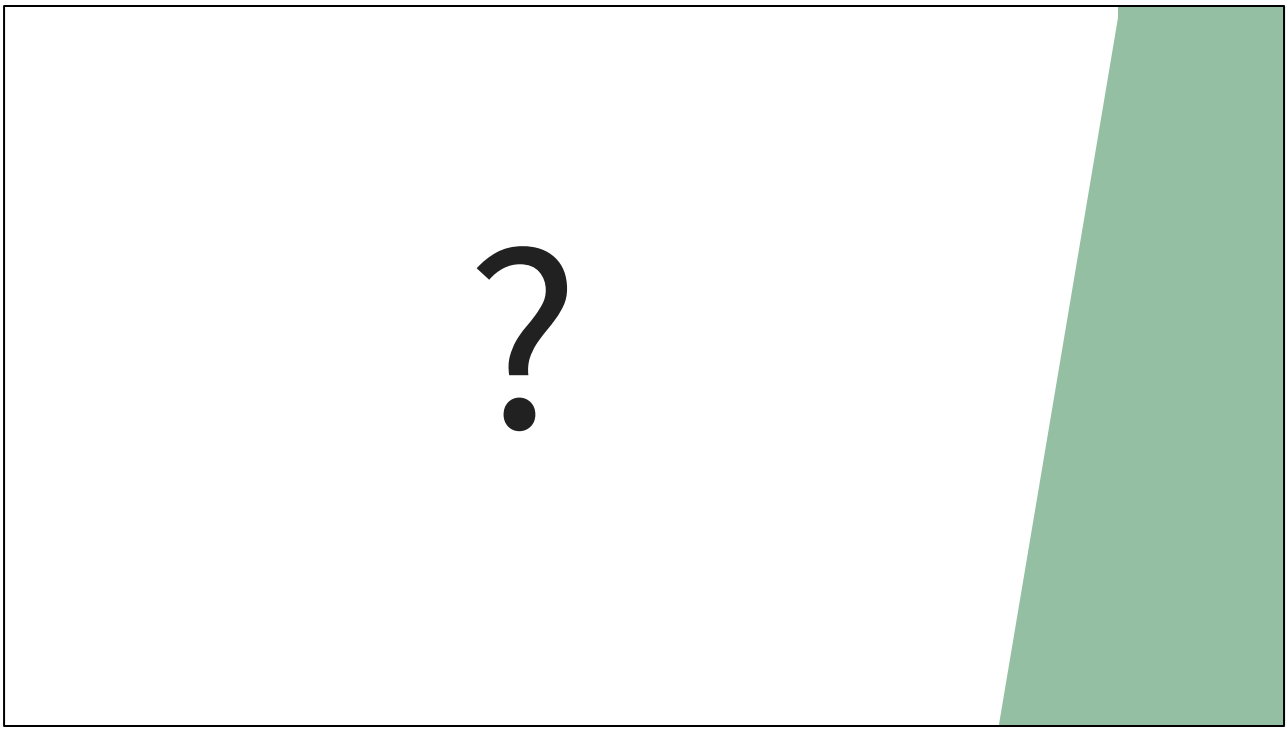
Oh the humanity!

So, Just like any protocol, it's only good so far as people are willing to enforce it. Google was one of the groups behind the sitemap protocol, but they do accept sitemaps in Google Search Console that don't follow the protocol, particularly around whether the sitemap is at the root of the section that it represents. There are humans behind every bot, and there are humans behind every website, so there is going to be messiness. Here are some of the things we've seen:

- Sitemaps located in obscure subfolders
- Sitemaps on one domain with URLs for different domains or subdomains
- Duplicate / Triplicate URLs
- URLs with spaces
- Staging URLs
- Relative URLs
- Local URLs
- URLs with ports declared
- URLs missing file extensions where they're needed
- URLs beginning `https:/`

Don't let this happen to you!

Do we have any questions about sitemaps?



Are there any questions?

Robots.txt

A robots.txt file signals

What to index and what not to index

How quickly or slowly bots should work in the site

So what is a robots.txt file? It's a signal to bots of what you want indexed, and what you don't want indexed, as well as of the posted speed limit for requests.

This file is NOT a setting that can actually control bots.

- Bots programmed with bad manners will ignore your requests and attempt to go all over your site, as fast as they can.

So you'll want to make good use of your firewall to shut things down if the requests go over your rate limit, and if you really need to keep bots out of particular areas of your site, you need to put that behind authentication. Reference: http://www.robotstxt.org/faq/blockjustbad.html

So why have one then? Bots programmed with good manners will follow your settings, and Google, Bing, Search.gov, and the Internet Archive definitely will. Because of this, you can mostly control what will appear in the Google index, and you can totally control what appears in Search.gov.

Robots Exclusion Protocol

```
https://www.agency.gov/robots.txt

<meta name="robots" content="noindex" />
```

The Robots Exclusion Protocol was first published in 1994, and there are two ways to use it. You can place a robots.txt file at the root of your domain, or you can put a robots meta tag in the head of a given page.

Robots Exclusion Protocol

```
User-agent:

Disallow:

# You can add comments
```

This is the entirety of the formal protocol - there are only two fields types. User-agent, specifying which bots you're targeting with particular commands, and Disallow, saying what you want the bots to stay out of.

Additional Fields

```
Crawl-delay:

Allow:

Sitemap:
```

There are a few other fields that have become standardized, though they're not part of the official standard.

Crawl-delay is your posted speed limit. The number here is the number of seconds ***in between*** requests by the bot.

Allow is the opposite of Disallow, and is helpful to allow a particular bot in a location where you want others to stay out, or to add an exception for a file in a folder that you've Disallowed.

Sitemap we talked about earlier.

Additional Fields

```
Crawl-delay: 10

Allow:

Sitemap:
```

```
  500,000 URLs
      x 10 seconds
5,000,000 seconds
```

58 days

A note about Crawl delay. This applies to all requests, whether the bot is actually crawling your site, or requesting pages on your sitemap - 10 is a the most common crawl delay I see, I think it's the default in some content management systems. If your site's really big, though, it can have a huge impact on the crawl of your site. It's basic math. If you've got 500,000 pages and a ten second crawl delay, it would take almost 58 days to get through the whole site just one time.

```
User-agent: usasearch

Crawl-delay: 2


User-agent: *

Crawl-delay: 10
```

So what we've been requesting is that sites let us work in their site more rapidly than the other bots. You can add different settings for different bots, like this.

Robots meta tag

noindex / index

nofollow / follow

```
<meta name="robots" content="noindex, nofollow" />

<meta name="robots" content="noindex, follow" />

<meta name="robots" content="index, nofollow" />
```

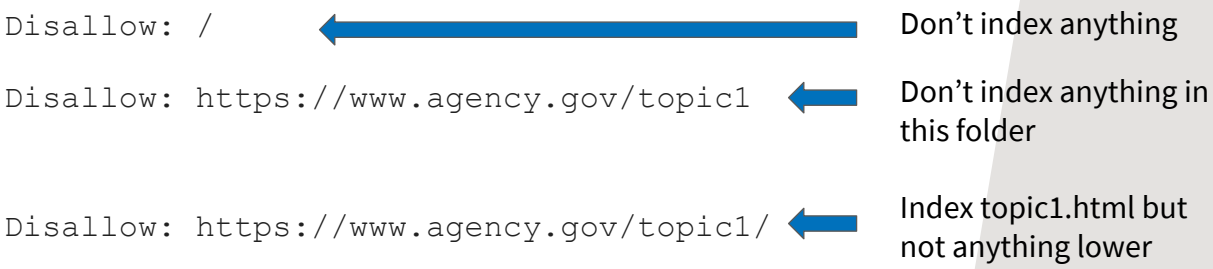
I mentioned that you can place robots tags at the page level, too. This is done through a meta tag in the head of the page, and you can tell a bot to not index the page or follow the links on the page. Or you could tell it to follow links but not index the page, or to index but not follow.

“Noindex, follow” is a really great setting to put on your content index pages - like for a given tag or list, where the items on the list would be good search results, but the list itself wouldn’t be that helpful as a result.

There are also some additional meta tags that Google pays attention to, and I’ve put a link in the slides to that info:
https://support.google.com/webmasters/answer/79812?hl=en

Robots Pitfalls

Mind the /



First, you need to be really careful with your slashes. They have different meanings at different levels.

Mind the /

```
Disallow: /
Allow:
```

Totally blocked

or

```
Disallow:
Allow: /
```

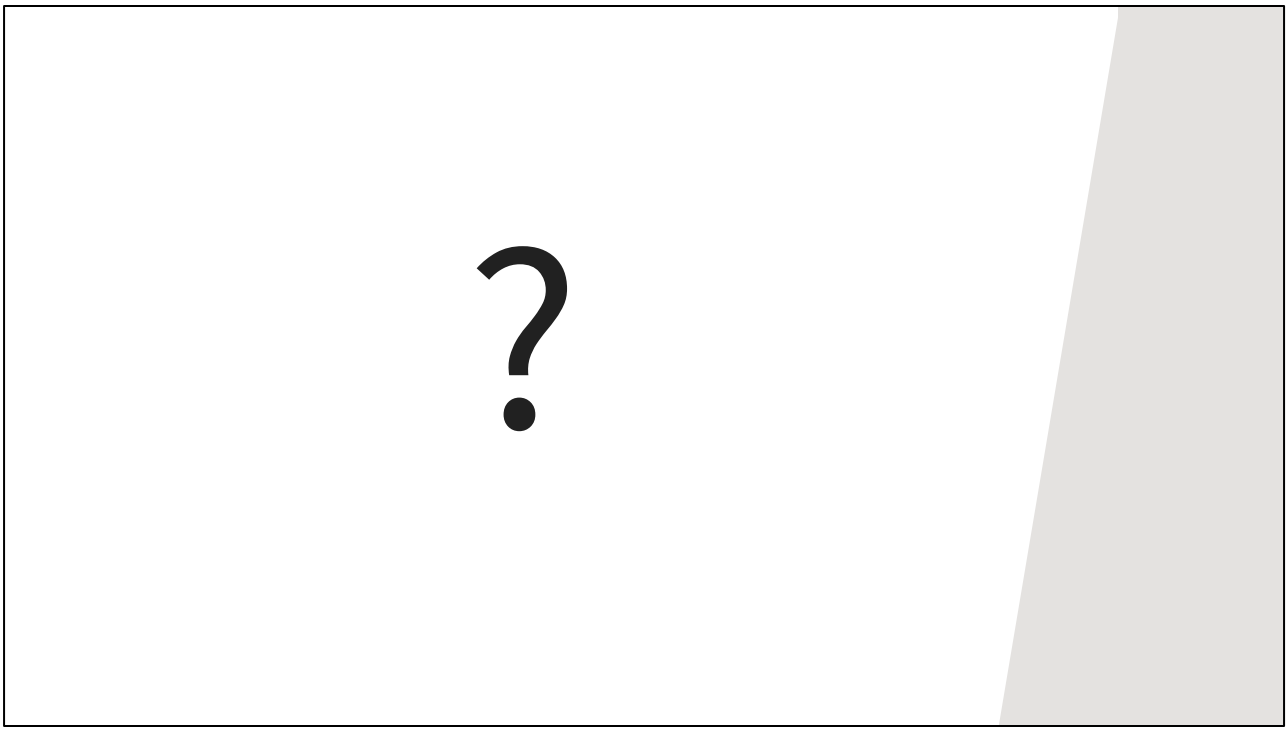
Totally open

Another way you need to be careful with your slashes is whether you're allowing or disallowing everything. Adding just a slash means "everything", and a blank means "nothing". So Disallow slash is the same as Allow blank, and vice versa.

More Pitfalls

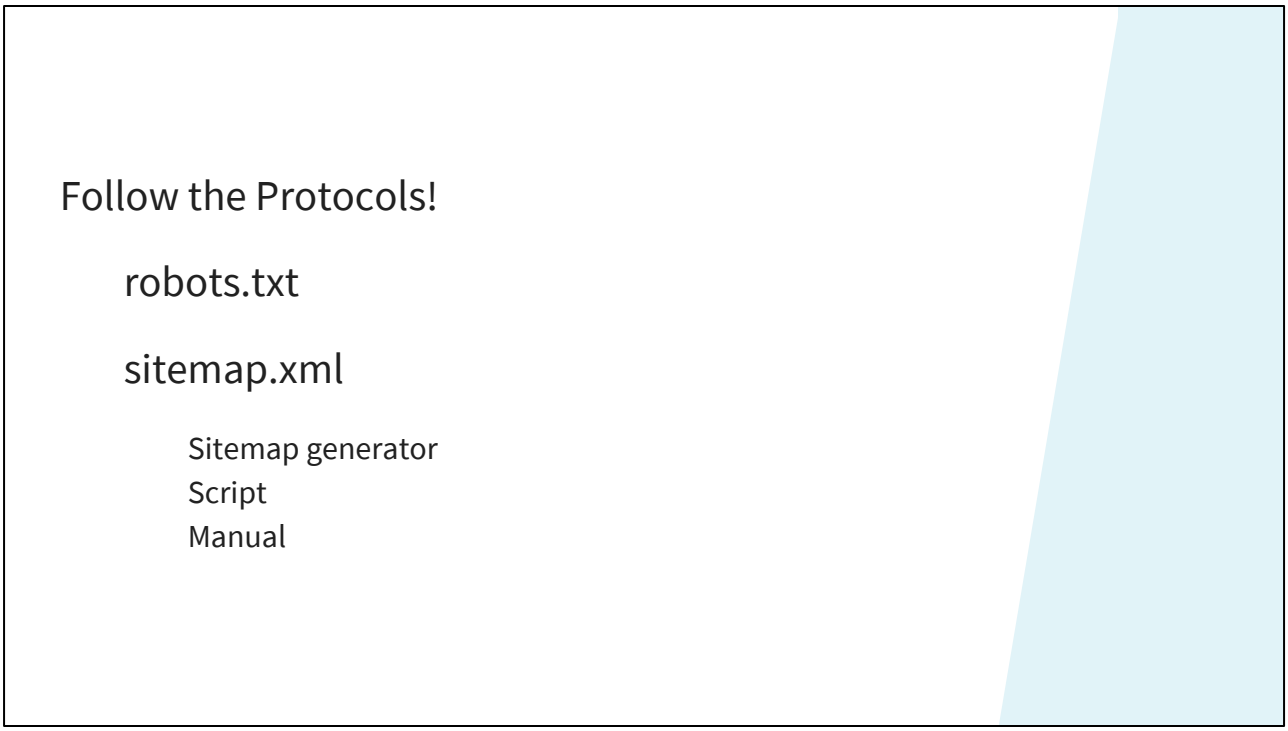
- Badly formatted lines
- Non-standard fields

And the last pitfalls I'll mention are getting sloppy or creative. As I mentioned, computers are dumb and won't be able to infer what you're talking about.



Have we gotten any more questions?

What about multiple publishing platforms?



You just want to follow the protocols.

Each system's subdomain will have its own robots.txt and sitemap.xml, or sitemap index.

If you use different publishing platforms, get as much automated as you can

- Use a sitemap generator
  - Drupal module
  - Yoast has a wordpress plugin and a drupal module
  - Static sites may have tools as well, such as the Jekyll-sitemap gem we use on GitHub Pages.

If you don't have a content management system, but you have a publication log, you might be able to write a script that would generate a sitemap when new entries appear in the log or another content inventory you might have.

And then if you have to, you could generate a sitemap manually.

- Most SEO tools will crawl your site to audit them and can generate an xml sitemap from what they find.
- Then you would post it to your web server
- Lather, rinse, repeast

Multiple platforms, same subdomain

Each system will produce its own sitemap

List all the sitemaps on the robots.txt

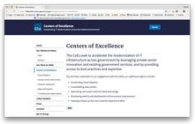
Some sites have multiple systems supporting the same subdomain. For instance, the main site has migrated to Drupal, but there are a bunch of folders that are still static and will be for a while.

- Each system should produce its own sitemap, and it would be placed at the root for that system. So you might have one at agency.gov/sitemap.xml, and another at agency.gov/subtopicA/sitemap.xml, and the last one at agency.gov/subtopicB/sitemap.xml
- And then you would list all of them in the robots.txt file for the domain, one per line.

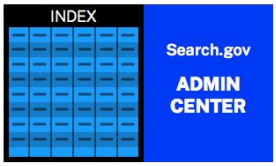
What's the relationship between sitemaps and search?

So, this question is why I wanted to do this session. I've gotten a lot of questions asking how a sitemap will control the search in Search.gov. As we've established, sitemaps support the indexing side of search engines, and then people querying the search engine is a separate process.

Let's go back to the pictures we were using earlier...

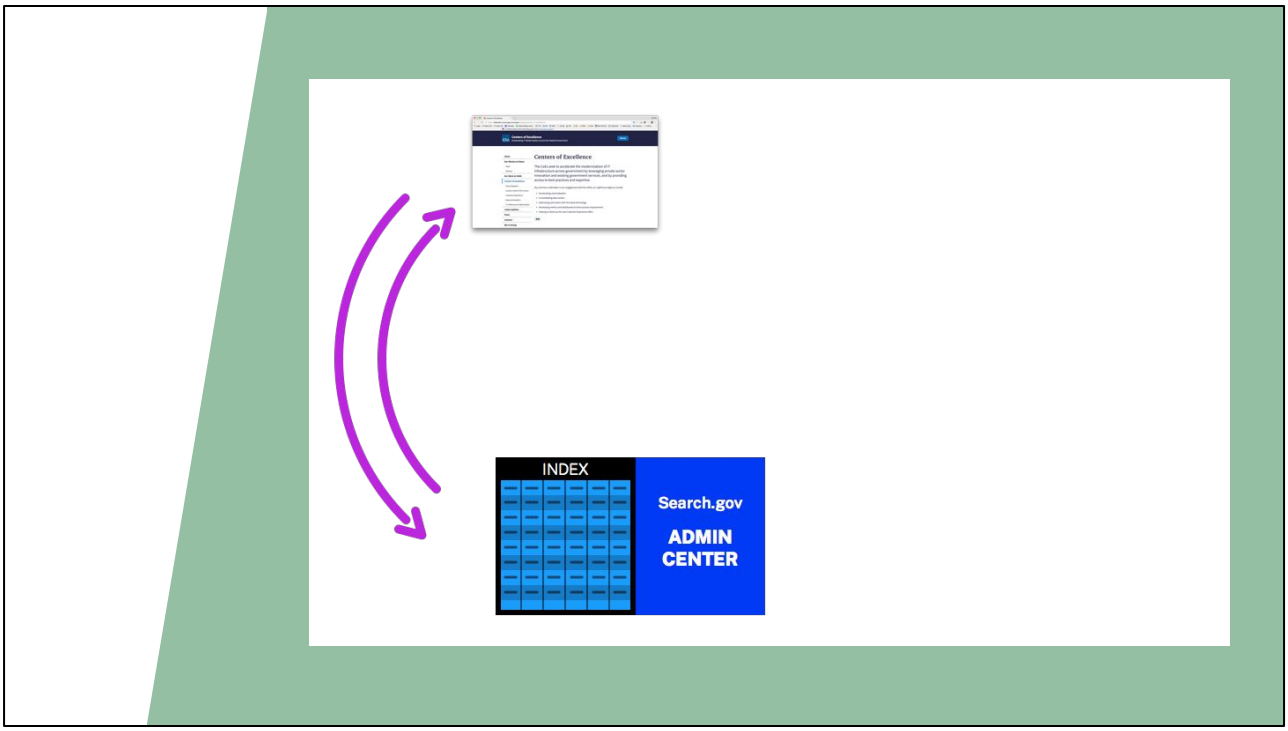


Government Website

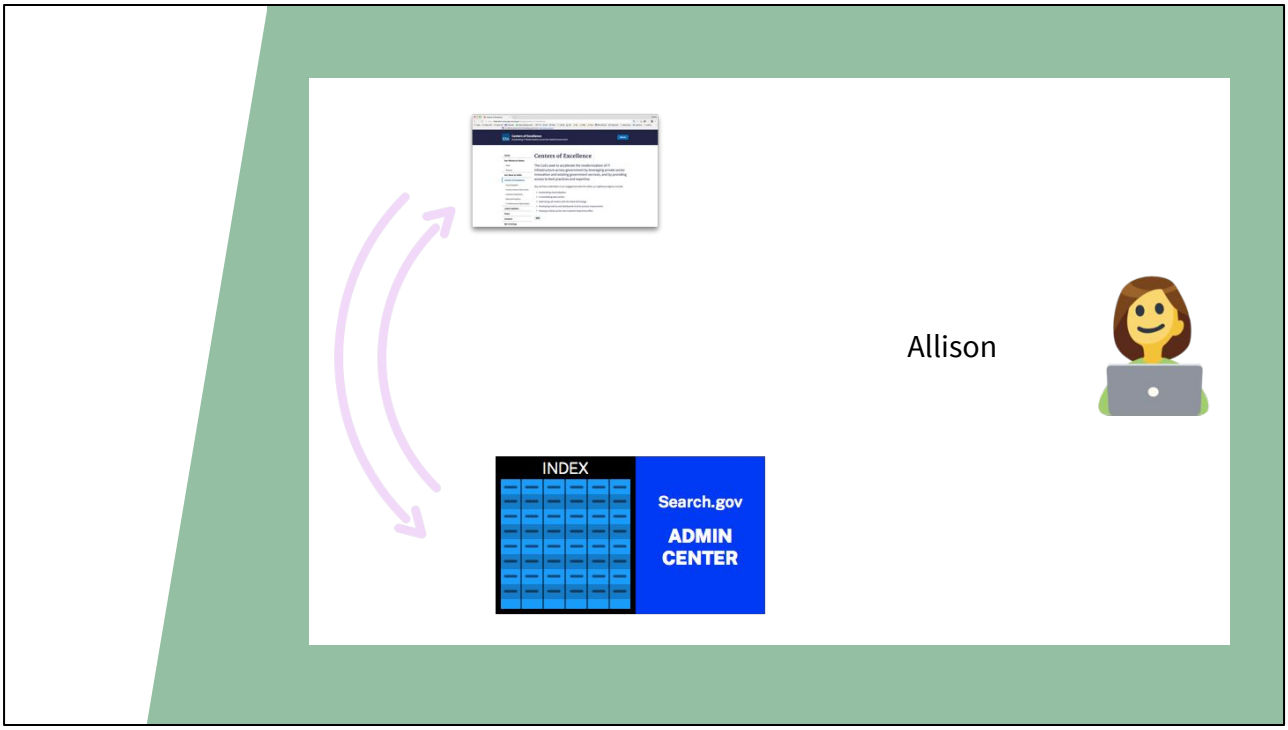


Search.gov

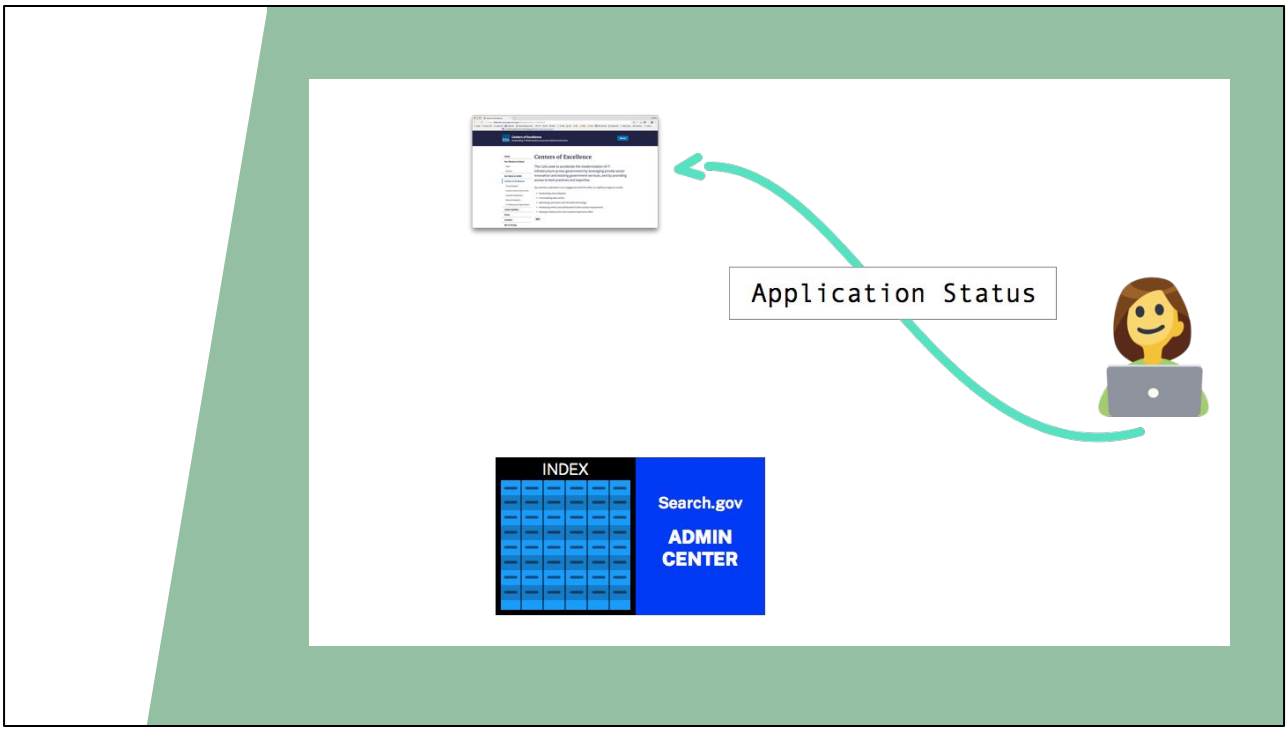
… Except now instead of Google, the search engine is Search.gov.



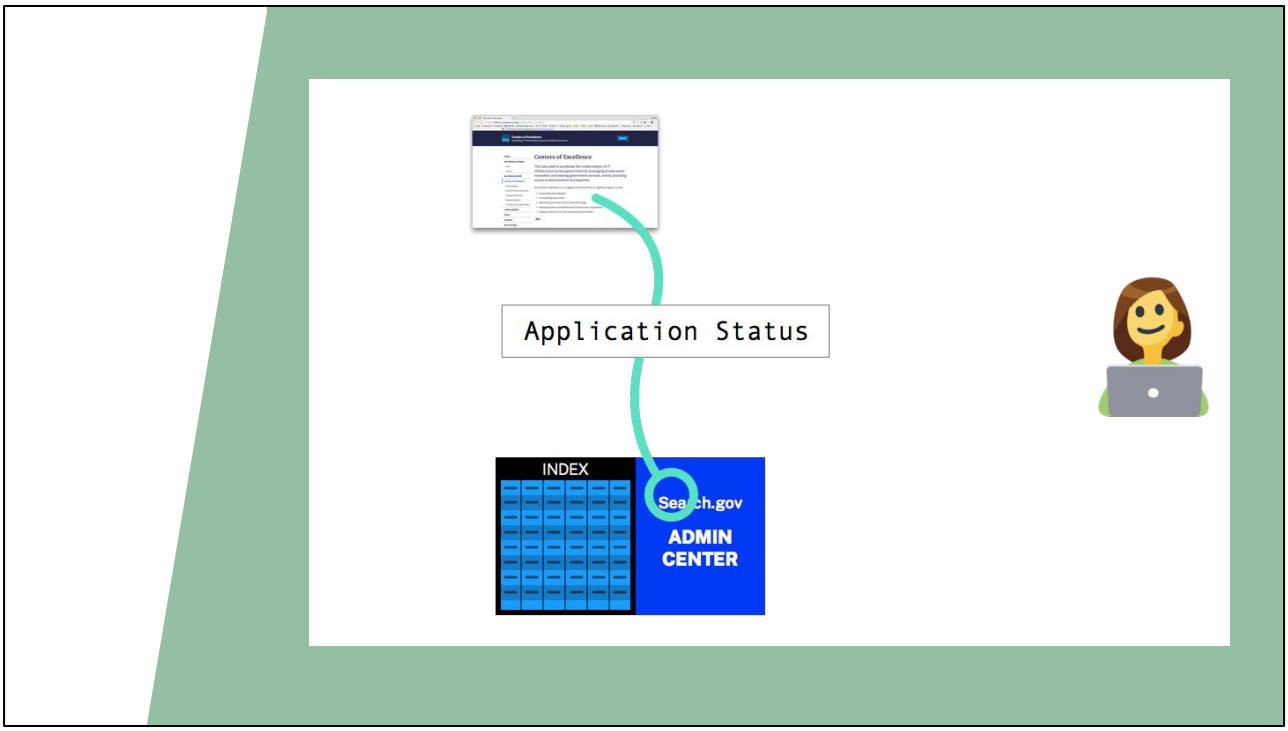
Just like Google, we will discover the URLs on your website, and then we'll pull data from them and stick it in our indexes along with other agencies content. We leverage your sitemaps, and honor your robots.txt settings.



Now here comes Allison.

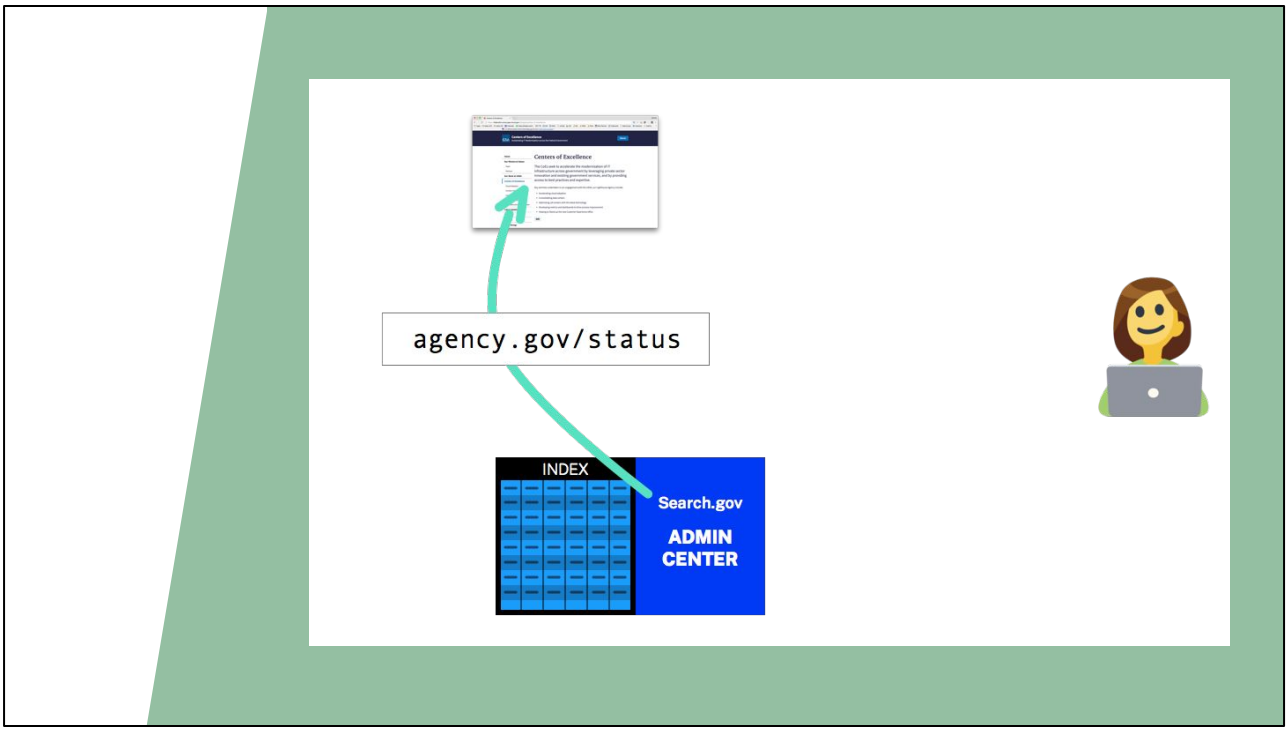


She wants to check on the status of her application. She knows what agency she's dealing with, so she goes right to their website and uses the search box to find the application status tool.



The difference between now and when she was searching google for passport application info, is that the agency whose site she's searching works with us, and has used our Admin Center to set up a search configuration that will target only the relevant content from our big index.

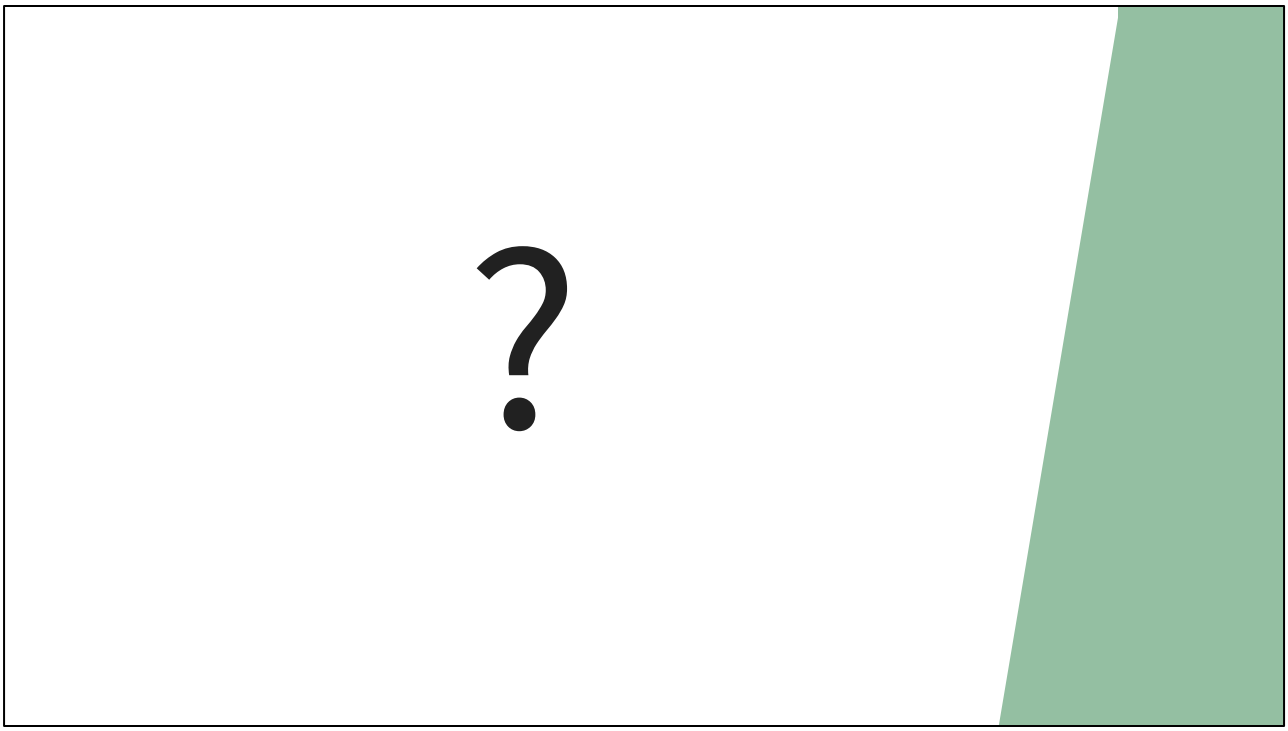
So the query is passed to us from the website, we process the query according to the search settings they've put in place, and we show her a page of results ...



She selects the status check tool, and is brought back to the agency's website to get her status.

So the relationship between sitemaps and your search configurations is minimal. The sitemaps support indexing, but the admin center settings control what exactly a person is going to be searching against when they use your search box.

That brings us to the end of the presentation...



Have any questions come in?

Search Engines discover URLs and parse data into their indexes to search against some other time.

XML Sitemaps & Robots.txt assist in URL discovery.

Indexing and search interact with the search engine from opposite directions.

So we covered a LOT today, we'll be circulating the slides with the links to all the references I mentioned, and of course when questions come up, just reach out

search@support.digitalgov.gov

202-505-5315

http://search.gov/manual/

Thank you!

Thank you very much!