

Duplicate Content in a Post-Panda World

In early 2011, Google launched the first phase of the “Panda” update, which would prove to be a wake-up call for SEO issues webmasters had been ignoring for too long. One of those issues was duplicate content. While duplicate content as an SEO problem has been around for years, the way Google handles it has evolved dramatically and seems to only get more complicated with every update. Panda upped the ante even more.

This document is an attempt to cover the topic of duplicate content, as it stands in 2011, in depth. This is designed to be a comprehensive resource – a complete discussion of what duplicate content is, how it happens, how to diagnose it, and how to fix it. Maybe we’ll even round up a few rogue pandas along the way.

I. What Is Duplicate Content?

Let’s start with the basics. Duplicate content exists when any two (or more) pages share the same content. If you’re a visual learner, here’s an illustration for you:



Easy enough, right? So, why does such a simple concept cause so much difficulty? One problem is that people often make the mistake of thinking that a “page” is a file or document sitting on their web server. To a crawler (like Googlebot), a page is any unique URL it happens to find, usually through internal or external links. Especially on large, dynamic sites, creating two URLs that land on the same content is surprisingly easy (and often unintentional).

II. Why Do Duplicates Matter?

Duplicate content as an SEO issue was around long before the Panda update, and has taken many forms as the algorithm has changed. Here’s a brief look at some major issues with duplicate content over the years...

The Supplemental Index

In the early days of Google, just indexing the web was a massive computational challenge. To deal with this challenge, some pages that were seen as duplicates or just very low quality were stored in a secondary index called the “supplemental” index. These pages automatically became 2nd-class citizens, from an SEO perspective, and lost any competitive ranking ability.

Around late 2006, Google integrated supplemental results back into the main index, but those results were still often filtered out. You know you’ve hit filtered results anytime you see this warning at the bottom of a Google SERP:

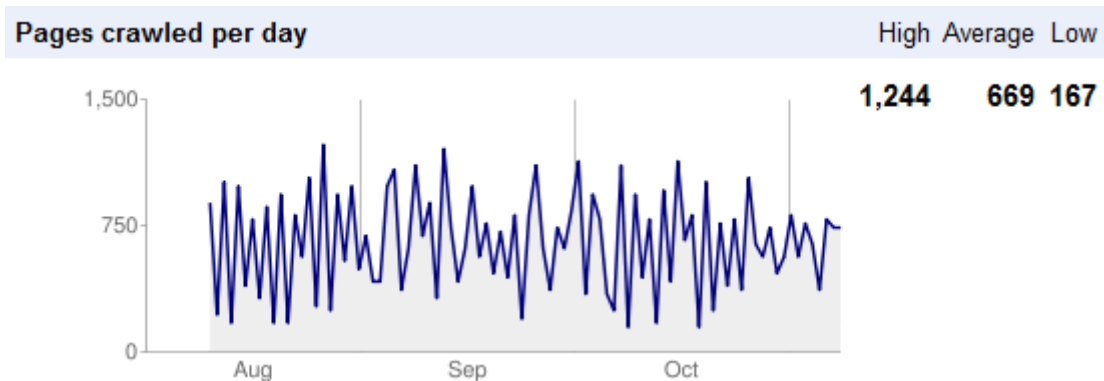
In order to show you the most relevant results, we have omitted some entries very similar to the 228 already displayed.
If you like, you can [repeat the search with the omitted results included](#).

Even though the index was unified, results were still “omitted”, with obvious consequences for SEO. Of course, in many cases, these pages really were duplicates or had very little search value, and the practical SEO impact was negligible, but not always.

The Crawl “Budget”

It’s always tough to talk limits when it comes to Google, because people want to hear an absolute number. There is no absolute crawl budget or fixed number of pages that Google will crawl on a site. There is, however, a point at which Google may give up crawling your site for a while, especially if you keep sending spiders down winding paths.

Although the “budget” isn’t absolute, even for a given site, you can get a sense of Google’s crawl allocation for your site in Google Webmaster Tools (under “Diagnostics” > “Crawl Stats”):



So, what happens when Google hits so many duplicate paths and pages that it gives up for the day? Practically, the pages you want indexed may not get crawled. At best, they probably won’t be crawled as often.

The Indexation “Cap”

Similarly, there’s no set “cap” to how many pages of a site Google will index. There does seem to be a [dynamic limit](#), though, and that limit is relative to the authority of the site. If you fill up your index with useless, duplicate pages, you may push out more important, deeper pages. For example, if you load up on 1000s of internal search results, Google may not index all of your product pages. Many people make the mistake of thinking that more indexed pages is better. I’ve seen too many situations where the opposite was true. All else being equal, bloated indexes dilute your ranking ability.

The Penalty Debate

Long before Panda, a debate would erupt every few months over whether or not there was a duplicate content penalty. While these debates raised valid points, they often focused on semantics – whether or not duplicate content caused a Capital-P Penalty. While I think the conceptual difference between penalties and filters matters, the upshot for a site owner is often the same. If a page isn’t ranking (or even indexed) because of duplicate content, then you’ve got a problem, no matter what you call it.

The Panda Update

Since Panda (starting in February 2011), the impact of duplicate content has become much more severe in some cases. It used to be that duplicate content could only harm that content itself. If you had a duplicate, it might go supplemental or get filtered out. Usually, that was ok. In extreme cases, a large number of duplicates could bloat your index or cause crawl problems and start impacting other pages.

Panda made duplicate content part of a broader quality equation – now, a duplicate content problem can impact the entire site. If you’re hit by Panda, non-duplicate pages may lose ranking power, stop ranking altogether, or even fall out of the index. Duplicate content is no longer an isolated problem.

III. Three Kinds of Duplicates

Before we dive into examples of duplicate content and the tools for dealing with them, I’d like to cover 3 broad categories of duplicates. They are: (1) True Duplicates, (2) Near Duplicates, and (3) Cross-domain Duplicates. I’ll be referencing these 3 main types in the examples later in the post.

(1) True Duplicates

A true duplicate is any page that is 100% identical (in content) to another page. These pages only differ by the URL:



(2) Near Duplicates

A near duplicate differs from another page (or pages) by a very small amount – it could be a block of text, an image, or even the order of the content:



An exact definition of “near” is tough to pin down, but I’ll discuss some examples in detail later.

(3) Cross-domain Duplicates

A cross-domain duplicate occurs when two websites share the same piece of content:



These duplicates could be either “true” or “near” duplicates. Contrary to what some people believe, cross-domain duplicates can be a problem even for legitimate, syndicated content.

IV. Tools for Fixing Duplicates

This may seem out of order, but I want to discuss the tools for dealing with duplicates before I dive into specific examples. That way, I can recommend the appropriate tools to fix each example without confusing anyone.

(1) 404 (Not Found)

Of course, the simplest way to deal with duplicate content is to just remove it and return a 404 error. If the content really has no value to visitors or search, and if it has no significant inbound links or traffic, then total removal is a perfectly valid option.

(2) 301 Redirect

Another way to remove a page is via a [301-redirect](#). Unlike a 404, the 301 tells visitors (humans and bots) that the page has permanently moved to another location. Human visitors seamlessly arrive at the new page. From an SEO perspective, most of the inbound link authority is also passed to the new page. If your duplicate content has a clear canonical URL, **and** the duplicate has traffic or inbound links, then a 301-redirect may be a good option.

(3) Robots.txt

Another option is to leave the duplicate content available for human visitors, but block it for search crawlers. The oldest and probably still easiest way to do this is with a [robots.txt file](#) (generally located in your root directory). It looks something like this:

```
User-agent: *
Disallow: /dupe-page.htm
Disallow: /dupe-folder/
```

One advantage of robots.txt is that it's relatively easy to block entire folders or even URL parameters. The disadvantage is that it's an extreme and sometimes unreliable solution. While robots.txt is effective for blocking uncrawled content, it's not great for removing content already in the index. The major search engines also seem to frown on its overuse, and don't generally recommend robots.txt for duplicate content.

(4) Meta Robots

You can also control the behavior of search bots at the page level, with a header-level directive known as the ["Meta Robots" tag](#) (or sometimes "Meta Noindex"). In its simplest form, the tag looks something like this:

```
<head>
  <meta name="ROBOTS" content="NOINDEX, NOFOLLOW" />
</head>
```

This directive tells search bots not to index this particular page or follow links on it. Anecdotally, I find it a bit more SEO-friendly than Robots.txt, and because the tag can be created dynamically with code, it can often be more flexible.

The other common variant for Meta Robots is the content value “NOINDEX, FOLLOW”, which allows bots to crawl the paths on the page without adding the page to the search index. This can be useful for pages like internal search results, where you may want to block certain variations (I’ll discuss this more later) but still follow the paths to product pages.

One quick note: there is no need to ever add a Meta Robots tag with “INDEX, FOLLOW” to a page. All pages are indexed and followed by default (unless blocked by other means).

(5) Rel=Canonical

In 2009, the search engines banded together to create the [Rel=Canonical directive](#), sometimes called just “Rel-canonical” or the “Canonical Tag”. This allows webmasters to specify a canonical version for any page. The tag goes in the page header (like Meta Robots), and a simple example looks like this:

```
<head>
  <link rel="canonical" href="http://www.example.com" />
</head>
```

When search engines arrive on a page with a canonical tag, they attribute the page to the canonical URL, regardless of the URL they used to reach the page. So, for example, if a bot reached the above page using the URL “www.example.com/index.html”, the search engine would **not** index the additional, non-canonical URL. Typically, it seems that inbound link-juice is also passed through the canonical tag.

It’s important to note that you need to clearly understand what the proper canonical page is for any given website template. Canonicalizing your entire site to just one page or the wrong pages [can be catastrophic](#).

(6) Google URL Removal

In [Google Webmaster Tools](#) (GWT), you can request that an individual page (or handful of pages) be manually removed from the index. Click on “Site configuration” > “Crawler access”, and you’ll see a series of 3 tabs. Click on the 3rd tab, “Remove URL”, to get this:



Test robots.txt Generate robots.txt **Remove URL**

New removal request

Enter the URL that you'd like to remove (case-sensitive) Removal Type

 Continue

Since this tool only removes one URL or path at a time and is completely at Google's discretion, it's usually a last-ditch approach to duplicate content. I just want to be thorough, though, and cover all of your options. Note: Google only removes these pages [for 90 days](#). For permanent removal, you need to 404, Robots.txt block or Meta Noindex the page. Removal via GWT is primarily a last defense when Google is being stubborn.

(7) Google Parameter Blocking

You can also use GWT to specify URL parameters that you want Google to ignore (which essentially blocks indexation of pages with those parameters). If you click on "Site Configuration" > "URL parameters", you'll get a list something like this:

Parameter	URLs monitored	What Googlebot should crawl	
utm_source	12	Let Googlebot decide (default)	Edit / Reset
utm_medium	11	Let Googlebot decide (default)	Edit / Reset
utm_campaign	7	Let Googlebot decide (default)	Edit / Reset
id	5	Let Googlebot decide (default)	Edit / Reset
text	5	Let Googlebot decide (default)	Edit / Reset

This list shows URL parameters that Google has detected, as well as the settings for how those parameters should be crawled. Keep in mind that the "Let Googlebot decide" setting doesn't reflect other blocking tactics, like Robots.txt or Meta Robots. If you click on "Edit", you'll get the following options:

Parameter: **id**

Does this parameter change page content seen by the user?

Save

Cancel

Google changed these recently, and I find the new version a bit confusing, but essentially "Yes" means the parameter is important and should be indexed, while "No" means the parameter indicates a duplicate. The GWT tool seems to be effective (and can be fast), but I don't usually recommend it as a first line of defense. It won't impact other search engines, and it can't be read by SEO tools and monitoring software. It could also be modified by Google at any time.

(8) Bing URL Removal

[Bing Webmaster Center](#) (BWC) has tools very similar to GWT's options above. Actually, I think the Bing parameter blocking tool came before Google's version. To request a URL removal in Bing, click on the "Index" tab and then "Block URLs" > "Block URL and Cache". You'll get a pop-up like this:



Block URL and Cache x

What would you like to block?

Page only

Directory

Entire site

Next

BWC actually gives you a wider range of options, including blocking a directory and your entire site. Obviously, that last one usually isn't a good idea.

(9) Bing Parameter Blocking

In the same section of BWC ("Index"), there's an option called "URL Normalization". The name implies Bing treats this more like canonicalization, but there's only one option – "ignore". Like Google, you get a list of auto-detected parameters and can add or modify them:

Parameter <small>△</small>	Status	Source	Date
ft	Disabled	Bing	11/7/2011
id	Disabled	Bing	11/7/2011
max	Disabled	Bing	11/7/2011
msg	Disabled	Bing	11/7/2011

As with the GWT tools, I'd consider the Bing versions to be a last resort. Generally, I'd only use these tools if other methods have failed, and one search engine is just giving you grief.

(10) Rel=Prev & Rel=Next

Just this year (September 2011), Google gave us a new tool for fighting a particular form of near-duplicate content – paginated search results. I’ll describe the problem in more detail in the next section, but essentially paginated results are any searches where the results are broken up into chunks, with each chunk (say, 10 results) having its own page/URL.

You can now tell Google how paginated content connects by using a pair of tags much like Rel-Canonical. They’re called [Rel-Prev and Rel-Next](#). Implementation is a bit tricky, but here’s a simple example:

```
<head>
  <link rel="prev" href="http://www.example.com/search/2" />
  <link rel="next" href="http://www.example.com/search/4" />
</head>
```

In this example, the search bot has landed on page 3 of search results, so you need two tags: (1) a Rel-Prev pointing to page 2, and (2) a Rel-Next pointing to page 4. Where it gets tricky is that you’re almost always going to have to generate these tags dynamically, as your search results are probably driven by one template.

While initial results suggest these tags do work, they’re not currently honored by Bing, and we really don’t have much data on their effectiveness. I’ll briefly discuss other methods for dealing with paginated content in the next section.

(11) Syndication-Source

In November of 2010, Google [introduced a set of tags](#) for publishers of syndicated content. The Meta Syndication-Source directive can be used to indicate the original source of a republished article, as follows:

```
<head>
  <meta name="syndication-source" content="http://example.com/news" />
</head>
```

Even Google’s own advice on when to use this tag and when to use a cross-domain canonical tag are a little bit unclear. Google launched this tag as “experimental”, and I’m not sure they’ve publicly announced a status change. It’s something to watch, but don’t rely on it.

(12) Internal Linking

It’s important to remember that your best tool for dealing with duplicate content is to not create it in the first place. Granted, that’s not always possible, but if you find yourself having to

patch dozens of problems, you may need to re-examine your internal linking structure and site architecture.

When you do correct a duplication problem, such as with a 301-redirect or the canonical tag, it's also important to make your other site cues reflect that change. It's amazing how often I see someone set a 301 or canonical to one version of a page, and then continue to link internally to the non-canonical version and fill their XML sitemap with non-canonical URLs. Internal links are strong signals, and sending mixed signals will only cause you problems.

(13) Don't Do Anything

Finally, you can let the search engines sort it out. This is what Google recommended you do for years, actually. Unfortunately, in my experience, especially for large sites, this is almost always a bad idea. It's important to note, though, that not all duplicate content is a disaster, and Google certainly can filter some of it out without huge consequences. If you only have a few isolated duplicates floating around, leaving them alone is a perfectly valid option.

V. Examples of Duplicate Content

So, now that we've worked backwards and sorted out the tools for fixing duplicate content, what does it actually look like in the wild? I'm going to cover a wide range of examples that represent the issues you can expect on a real website. Throughout this section, I'll reference the solutions listed in Section IV – for example, a reference to a 301-redirect will cite (IV-2).

(1) "www" vs. Non-www

For sitewide duplicate content, this is probably the biggest culprit. Whether you've got bad internal paths or have attracted links and social mentions to the wrong URL, you've got both the "www" version and non-www (root domain) version of your URLs indexed:



Most of the time, a 301-redirect (IV-2) is your best choice here. This is a common problem, and Google is good about honoring redirects for cases like these.

You may also want to set your preferred address in Google Webmaster Tools. Under "Site Configuration" > "Settings", you should see a section called "Preferred domain":

- Preferred domain**
- Don't set a preferred domain
 - Display URLs as **www.30go30.com**
 - Display URLs as **30go30.com**

There's a quirk in GWT where, to set a preferred domain, you may have to create GWT profiles for both your "www" and non-www versions of the site. While this is annoying, it won't cause any harm. If you're having major canonicalization issues, I'd recommend it. If you're not, then you can leave well enough alone and let Google determine the preferred domain.

(2) Staging Servers

While much less common than (1), this problem is often also caused by subdomains. In a typical scenario, you're working on a new site design for a relaunch, your dev team sets up a subdomain with the new site, and they accidentally leave it open to crawlers. What you end up with is two sets of indexed URLs that look something like this:



Your best bet is to prevent this problem before it happens, by blocking the staging site with Robots.txt (IV-3). If you find your staging site indexed, though, you'll probably need to 301-redirect (IV-2) those pages or Meta Noindex them (IV-4).

(3) Trailing Slashes ("/")

This is a problem people often have questions about, although it's less of an SEO issue than it once was. Technically, in the original HTTP protocol, a URL with a trailing slash and one without it were different URLs. Here's a simple example:



These days, almost all browsers automatically add the trailing slash behind the scenes and resolve both versions the same way. Matt Cutts did a recent video suggesting that Google [automatically canonicalizes](#) these URLs in "the vast majority of cases".

(4) Secure (https) Pages

If your site has secure pages (designated by the "https:" protocol), you may find that both secure and non-secure versions are getting indexed. This most frequently happens when navigation links from secure pages – like shopping cart pages – also end up secured, usually due to relative paths, creating variants like this:

 www.example.com

 https://www.example.com

Ideally, these problems are solved by the site-architecture itself. In many cases, it's best to Noindex (IV-4) secure pages – shopping cart and check-out pages have no place in the search index. After the fact, though, your best option is a 301-redirect (IV-2). Be cautious with any sitewide solutions – if you 301-redirect all “https:” pages to their “http:” versions, you could end up removing security entirely. This is a tricky problem to solve and should be handled carefully.

(5) Home-page Duplicates

While problems (1)-(3) can all create home-page duplicates, the home-page has a couple unique problems of its own. The most typical problem is that both the root domain and the actual home-page document name get indexed. For example:

 www.example.com

 www.example.com/index.htm

Although this problem can be solved with a 301-redirect (IV-2), it's often a good idea to put a canonical tag on your home-page (IV-5). Home pages are uniquely afflicted by duplicates, and a proactive canonical tag can prevent a lot of problems.

Of course, it's important to also be consistent with your internal paths (IV-12). If you want the root version of the URL to be canonical, but then link to “/index.htm” in your navigation, you're sending mixed signals to Google every time the crawlers visit.

(6) Session IDs

Some websites (especially e-commerce platforms) tag each new visitor with a tracking parameter. On occasion, that parameter ends up in the URL and gets indexed, creating something like this:

 www.example.com

 www.example.com/?session=12345678

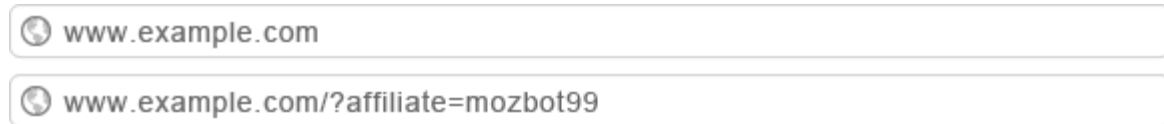
That image really doesn't do the problem justice, because in reality you can end up with a duplicate for virtually every visitor to your site, multiplied across all of the pages of your site. Session IDs in the URL can easily add 1000s of duplicate pages to your index.

The best option, if possible on your site/platform, is to remove the session ID from the URL altogether and store it in a cookie. There are very few good reasons to create these URLs. If

that's not feasible, implementing the canonical tag (IV-5) sitewide is a good bet. If you really get stuck, you can block the parameter in Google Webmaster Tools (IV-7) and Bing Webmaster Central (IV-9).

(7) Affiliate Tracking

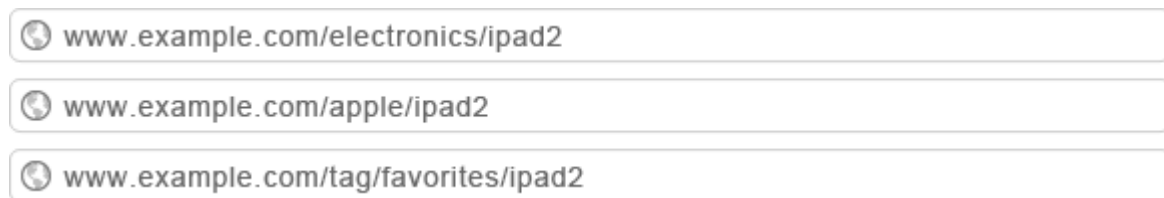
This problem looks a lot like (5) and happens when sites provide a tracking variable to their affiliates. This variable is typically appended to landing page URLs, like so:



The damage is usually a bit less extreme than (5), but it can still cause large-scale duplication. The solutions are similar to session IDs. Ideally, you can capture the affiliate ID in a cookie and 301-redirect (IV-3) to the canonical version of the page. Otherwise, you'll probably either need to use canonical tags (IV-5) or block the affiliate URL parameter.

(8) Duplicate Paths

Having duplicate paths to a page is perfectly fine, but when duplicate paths generate duplicate URLs, then you've got a problem. Let's say a product page can be reached one of 3 ways:



Here, the iPad2 product page can be reached by 2 categories and a user-generated tag. User-generated tags are especially problematic, because they can theoretically spawn unlimited versions of a page.

Ideally, these paths shouldn't be created at all. However a page is navigated to, it should only have one URL for SEO purposes. Some will argue that including navigation paths in the URL is a positive cue for site visitors, but even as someone with a usability background, I think the cons almost always outweigh the pros here.

If you already have variations indexed, then a 301-redirect (IV-3) or canonical tag (IV-5) are probably your best options. In many cases, implementing the canonical tag will be easier, since there may be too many variations to easily redirect. Long-term, though, you'll need to re-evaluate your site architecture.

(9) Functional Parameters

Functional parameters are URL parameters that change a page slightly but have no value for search and are essentially duplicates. For example, let's say that all of your product pages have a printable version, and that version has its own URL:

 www.example.com/product.php?id=1234

 www.example.com/product.php?id=1234&print=1

Here, the “print=1” URL variable indicates a printable version, which normally would have the same content but a modified template. Your best bet is to not index these at all, with something like a Meta Noindex (IV-4), but you could also use a canonical tag (IV-5) to consolidate these pages.

(10) International Duplicates

These duplicates occur when you have content for different countries which share the same language, all hosted on the same root domain (it could be subfolders or subdomains). For example, you may have an English version of your product pages for the US, UK, and Australia:

 www.example.com/us/product/ipad2

 www.example.com/uk/product/ipad2

 www.example.com/au/product/ipad2


Unfortunately, this one's a bit tough – in some cases, Google will handle it perfectly well and rank the appropriate content in the appropriate countries. In other cases, even with proper geo-targeting, they won't. It's often better to target the language itself than the country, but there are legitimate reasons to split off country-specific content, such as pricing.

If your international content does get treated as duplicate content, there's no easy answer. If you 301-redirect, you lose the page for visitors. If you use the canonical tag, then Google will only rank one version of the page. The “right” solution can be highly situational and really depends on the risk-reward tradeoff (and the scope of the filter/penalty).

(11) Search Sorts

So far, all of the examples I've given have been true duplicates. I'd like to dive into a few examples of “near” duplicates, since that concept is a bit fuzzy. A few common examples pop up with internal search engines, which tend to spin off many variants – sortable results, filters, and paginated results being the most frequent problems.

Search sort duplicates pop up whenever a sort (ascending/descending) creates a separate URL. While the two sorted results are technically different pages, they add no additional value to the search index and contain the same content, just in a different order. URLs might look like:


 www.example.com/search.php?keyword=ipad

 www.example.com/search.php?keyword=ipad&sort=desc

In most cases, it's best just to block the sortable versions completely, usually by adding a Meta Noindex (IV-4) selectively to pages called with that parameter. In a pinch, you could block the sort parameter in Google Webmaster Tools (IV-7) and Bing Webmaster Central (IV-9).

(12) Search Filters

Search filters are used to narrow an internal search – it could be price, color, features, etc. Filters are very common on e-commerce sites that sell a wide variety of products. Search filter URLs look a lot like search sorts, in many cases:

 www.example.com/search.php?category=laptop

 www.example.com/search.php?category=laptop?price=1000

The solution here is similar to (10) – don't index the filters. As long as Google has a clear path to products, indexing every variant usually causes more harm than good.

(13) Search Pagination

Pagination is an easy problem to describe and an incredibly difficult one to solve. Any time you split internal search results into separate pages, you have paginated content. The URLs are easy enough to visualize:

 www.example.com/search.php?category=laptop

 www.example.com/search.php?category=laptop?page=2

Of course, over 100s of results, one search can easily spin out dozens of near duplicates. While the results themselves differ, many important features of the pages (Titles, Meta Descriptions, Headers, copy, template, etc.) are identical. Add to that the problem that Google isn't a big fan of "search within search" (having their search pages land on yours).

In the past, Google has said to let them sort pagination out – problem is, they haven't done it very well. Recently, Google introduced Rel=Prev and Rel=Next (IV-10). Initial data suggests these tags work, but we don't have much data, they're difficult to implement, and Bing doesn't currently support them.

You have 3 other, viable options (in my opinion), although how and when they're viable depends a lot on the situation:

- (1) You can Meta Noindex, Follow pages 2+ of search results. Let Google crawl the paginated content but don't let them index it.
- (2) You can create a "View All" page that links to all search results at one URL, and let Google auto-detect it. This seems to be Google's other preferred option.
- (3) You can create a "View All" page and set the canonical tag of paginated results back to that page. This is unofficially endorsed, but the pages aren't really duplicates in the traditional sense, so some claim it violates the intent of Rel-canonical.

Adam Audette has a recent, in-depth [discussion of search pagination](#) that I highly recommend. Pagination for SEO is a very difficult topic and well beyond the scope of this post.

(14) Product Variations

Product variant pages are pages that branch off from the main product page and only differ by one feature or option. For example, you might have a page for each color a product comes in:



It can be tempting to want to index every color variation, hoping it pops up in search results, but in most cases I think the cons outweigh the pros. If you have a handful of product variations and are talking about dozens of pages, fine. If product variations spin out into 100s or 1000s, though, it's best to consolidate. Although these pages aren't technically true duplicates, I think it's ok to Rel-canonical (IV-5) the options back up to the main product page.

One site note: I purposely used "static" URLs in this example to demonstrate a point. Just because a URL doesn't have parameters, that doesn't make it immune to duplication. Static URLs (parameter-free) may look prettier, but they can be duplicates just as easily as dynamic URLs.

(15) Geo-keyword Variations

Once upon a time, "local SEO" meant just copying all of your pages 100s of times, adding a city name to the URL, and swapping out that city in the page copy. It created URLs like these:

 www.example.com/product/ipad2/new-york

 www.example.com/product/ipad2/chicago

 www.example.com/product/ipad2/miami

In 2011, not only is local SEO a lot more sophisticated, but these pages are almost always going to look like near-duplicates. If you have any chance of ranking, you're going to need to invest in legitimate, unique content for every geographic region you spin out. If you aren't willing to make that investment, then don't create the pages. They'll probably backfire.

(16) Other "Thin" Content

This isn't really an example, but I wanted to stop and explain a word we throw around a lot when it comes to content: "thin". While thin content can mean a variety of things, I think many examples of thin content are near-duplicates like (14) above. Whenever you have pages that vary by only a tiny percentage of content, you risk those pages looking low-value to Google. If those pages are heavy on ads (with more ads than unique content), you're at even more risk. When too much of your site is thin, it's time to revisit your content strategy.

(17) Syndicated Content

These last 3 examples all relate to cross-domain content. Here, the URLs don't really matter – they could be wildly different. Examples (16) and (17) only differ by intent. Syndicated content is any content you use with permission from another site. However you retrieve and integrate it, that content is available on another site (and, often, many sites).

While syndication is legitimate, it's still likely that one or more copies will get filtered out of search results. You could roll the dice and see what happens (IV-13), but conventional SEO wisdom says that you should link back to the source and probably set up a cross-domain canonical tag (IV-5). A cross-domain canonical looks just like a regular canonical, but with a reference to someone else's domain.

Of course, a cross-domain canonical tag means that, assuming Google honors the tag, your page won't get indexed or rank. In some cases, that's fine – you're using the content for its value to visitors. Practically, I think it depends on the scope. If you occasionally syndicate content to beef up your own offerings but also have plenty of unique material, then link back and leave it alone. If a larger part of your site is syndicated content, then you could find yourself running into trouble. Unfortunately, the canonical tag (IV-5) won't solve that problem, but it could keep you from getting penalized or having Panda-related problems.

(18) Scraped Content

Scraped content is just like syndicated content, except that you didn't ask permission (and might even be breaking the law). The best solution: QUIT BREAKING THE LAW!

Seriously, no de-duping solution is going to satisfy the scrapers among you, because most solutions will knock your content out of ranking contention. The best you can do is pad the scraped content with as much of your own, unique content as possible.

(19) Cross-ccTLD Duplicates

Finally, it's possible to run into trouble when you copy same-language content across countries – see example (9) above – even with separate Top-Level Domains (TLDs). Fortunately, this problem is fairly rare, but we see it with English-language content and even with some European languages. For example, I frequently see questions about Dutch Content on Dutch and Belgian domains ranking improperly.

Unfortunately, there's no easy answer here, and most of the solutions aren't traditional duplicate-content approaches. In most cases, you need to work on your targeting factors and clearly showing Google that the domain is tied to the country in question.

VI. Which URL Is Canonical?

I'd like to take a quick detour to discuss an important question – whether you use a 301-redirect or a canonical tag, how do you know which URL is actually canonical? I often see people making a mistake like this:

```
<head>
  <link rel="canonical" href="http://www.example.com/product.php">
</head>
```

The problem is that “product.php” is just a template – you've now collapsed all of your products down to a single page (that probably doesn't even display a product). In this case the canonical version probably includes a parameter, like “id=1234”.

The canonical page isn't always the simplest version of the URL – it's the simplest version of the URL that generates UNIQUE content. Let's say you have these 3 URLs that all generate the same product page:

www.example.com/product.php?id=1234

www.example.com/product.php?id=1234&print=1

www.example.com/product.php?id=1234&session=12345678

Two of these versions are essentially duplicates, and the “print” and “session” parameters represent variations on the main product page that should be de-duped. The “id” parameter is essential to the content, though – it determines which product is actually being displayed.

So, consider yourself warned. As much trouble as rampant duplicates can be, bad canonicalization can cause even more damage in some cases. Plan carefully, and make absolutely sure you select the correct canonical versions of your pages before consolidating them.

VII. Tools for Diagnosing Duplicates

So, now that you recognize what duplicate content looks like, how do you go about finding it on your own site? Here are a few tools to get you started – I won’t claim it’s a complete list, but it covers the bases:

(1) Google Webmaster Tools

In Google Webmaster Tools, you can pull up a list of duplicate TITLE tags and Meta Descriptions Google has crawled. While these don’t tell the whole story, they’re a good starting point. Many URL-based duplicates will naturally generate identical Meta data. In your GWT account, go to “Diagnostics” > “HTML Suggestions”, and you’ll see a table like this:

Meta description	Pages
Duplicate meta descriptions	149
Long meta descriptions	0
Short meta descriptions	0

Title tag	Pages
Missing title tags	0
Duplicate title tags	73
Long title tags	0

You can click on “Duplicate meta descriptions” and “Duplicate title tags” to pull up a list of the duplicates. This is a great first stop for finding your trouble-spots.

(2) Google’s Site: Command

When you already have a sense of where you might be running into trouble and need to take a deeper dive, Google’s “site:” command is a very powerful and flexible tool. What really makes “site:” powerful is that you can use it in conjunction with other search operators.

Let’s say, for example, that you’re worried about home-page duplicates. To find out if Google has indexed any copies of your home-page, you could use the “site:” command with the “intitle:” operator, like this:



Put the title in quotes to capture the full phrase, and always use the root domain (leave off “www”) when making a wide sweep for duplicate content. This will detect both “www” and non-www versions.

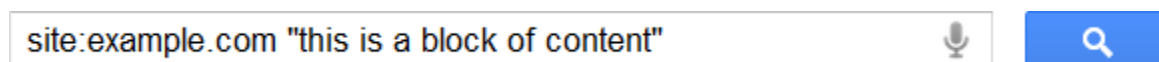
Another powerful combination is “site:” plus the “inurl:” operator. You could use this to detect parameters, such as the search-sort problem mentioned above:



The “inurl:” operator can also detect the protocol used, which is handy for finding out whether any secure (https:) copies of your pages have been indexed:



You can also combine the “site:” operator with regular search text, to find near-duplicates (such as blocks of repeated content). To search for a block of content across your site, just include it in quotes:

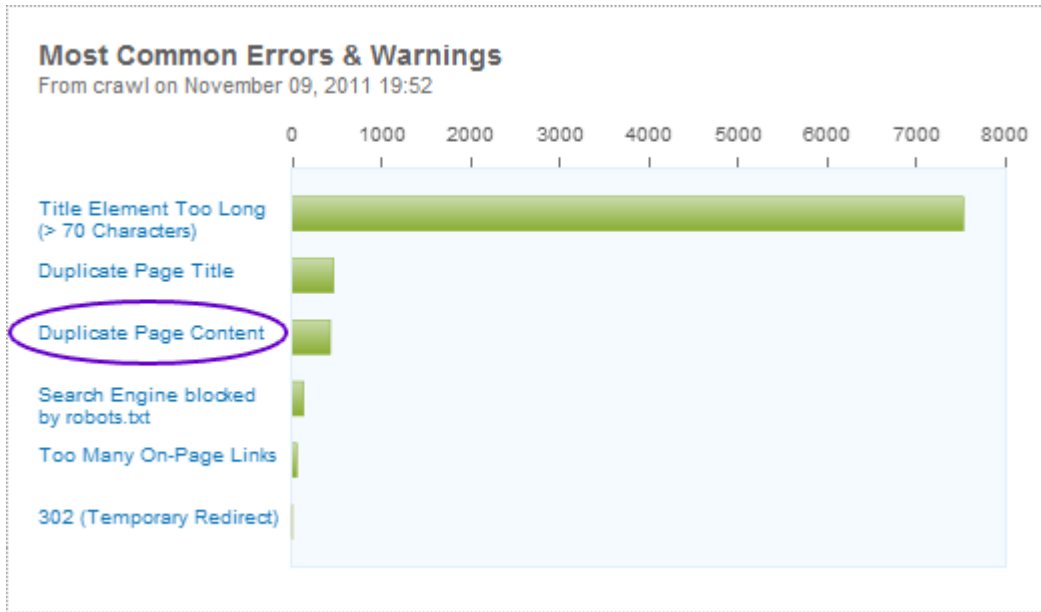


I should also mention that searching for a unique block of content in quotes is a cheap and easy way to find out if people have been scraping your site. Just leave off the “site:” operator and search for a long or unique block entirely in quotes.

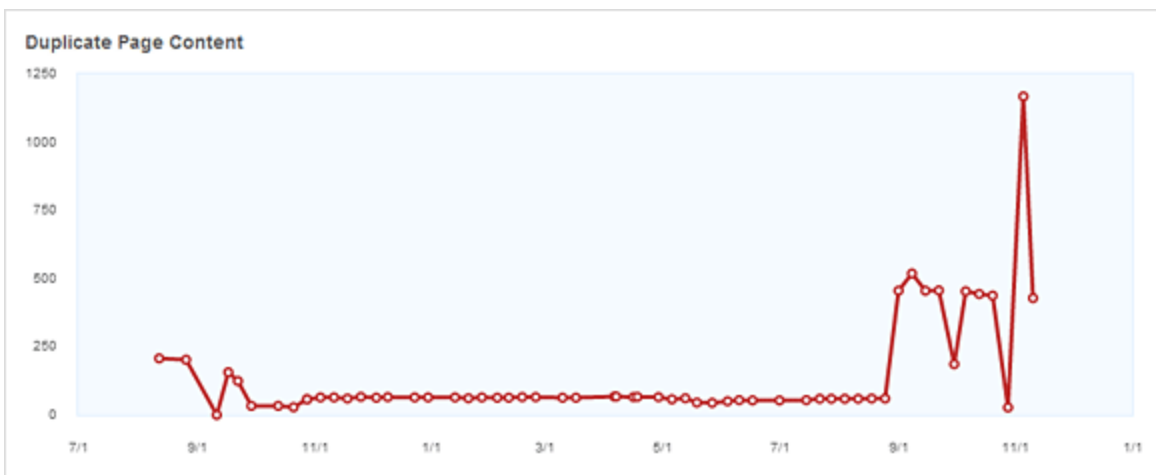
Of course, these are just a few examples, but if you really need to dig deep, these simple tools can be used in powerful ways. Ultimately, the best way to tell if you have a duplicate content problem is to see what Google sees.

(3) SEOmoz Campaign Manager

If you're an [SEOmoz PRO member](#), you have access to some additional tools for spotting duplicates in your Campaigns. In addition to duplicate page titles, the Campaign manager will detect duplicate content on the pages themselves. You can see duplicate pages we've detected from the Campaign Overview screen:



Click on the "Duplicate Page Content" link and you'll not only see a list of potential duplicates, but you'll get a graph of how your duplicate count has changed over time:



The historical graph can be very useful for determining if any recent changes you've made have created (or resolved) duplicate content issues.

Just a technical note, since it comes up a lot in Q&A – Our system currently uses a threshold of 95% to determine whether content is duplicated. This is based on the source code (not the text copy), so the amount of actual duplicate content may vary depending on the code/content ratio.

(4) Your Own Brain

Finally, it's important to remember to use your own brain. Finding duplicate content often requires some detective work, and over-relying on tools can leave some gaps in what you find. One critical step is to systematically navigate your site to find where duplicates are being created. For example, does your internal search have sorts and filters? Do those sorts and filters get translated into URL variables, and are they crawlable? If they are, you can use the "site:" command to dig deeper. Even finding a handful of trouble spots using your own sleuthing skills can end up revealing 1000s of duplicate pages, in my experience.

I Hope That Covers It

If you've made it this far: congratulations – you're probably as exhausted as I am. I hope that covers everything you'd want to know about the state of duplicate content in 2011, but if not, I'd be happy to answer questions in the comments. Dissenting opinions are welcome, too. Some of these topics, like pagination, are extremely tricky in practice, and there's often not one "right" answer.

This document was adapted from a blog post titled "[Duplicate Content in a Post-Panda World](#)", authored by Dr. Peter J. Meyers and published on the [SEOmoz blog](#) on November 16, 2011.