

Search



OMNI MARKETING INTERACTIVE  
Designing for people who search™


## Solutions & Best Practices for Managing Duplicate Content Delivery

By  
Shari Thurow, Founder & SEO Director

Copyright 1995-2019. All rights reserved.


### Workshop agenda:

- Identifying duplicate content
- What is duplicate content?
- Why should we care?
- How to determine duplicate content
- Duplicate content solutions
- Questions & answers



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



OMNI MARKETING INTERACTIVE

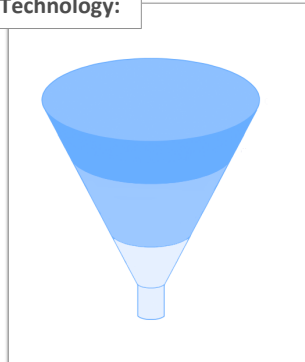
## Main goal of this workshop – mental model:

Human users:



Managing Duplicate Content Delivery

Technology:



Copyright 1995-2019. All rights reserved.



*If you check out some of the sample web pages used in this presentation, they are likely to look different.*

*The principles & guidelines that these screenshots illustrate are relevant long after a site has changed.*

Copyright ©1995-2019. All rights reserved.



## Solutions &amp; Best Practices for Managing Duplicate Content Delivery

**IDENTIFYING DUPLICATE CONTENT**

Copyright 1995-2019. All rights reserved.

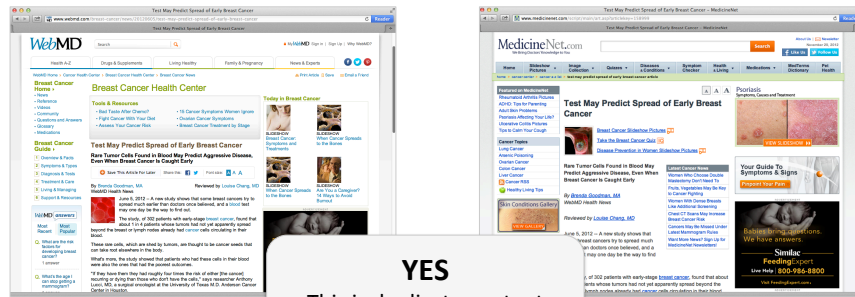


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

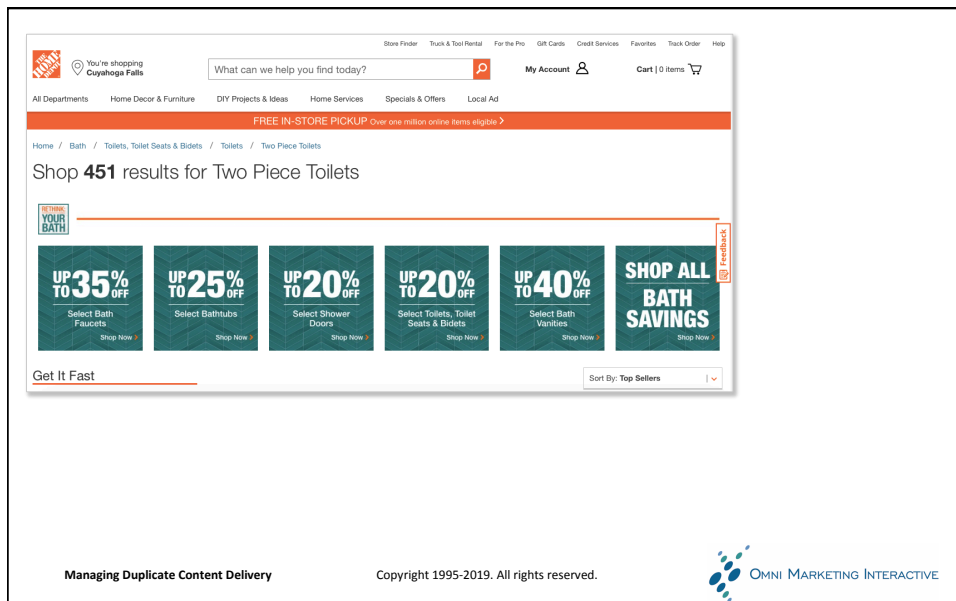


## Is this duplicate content?



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



Home Depot

You're shopping Cuyahoga Falls

What can we help you find today?

My Account

Cart | 0 items

Store Finder | Truck & Tool Rental | For the Pro | Gift Cards | Credit Services | Favorites | Track Order | Help

All Departments | Home Decor & Furniture | DIY Projects & Ideas | Home Services | Specials & Offers | Local Ad

FREE IN-STORE PICKUP Over one million online items eligible

Home / Bath / Toilets, Toilet Seats & Bidets / Toilets / Two Piece Toilets

Shop 451 results for Two Piece Toilets

Get It Fast

UP TO 35% OFF Select Bath Faucets

UP TO 25% OFF Select Bathroom

Get It Fast

Sort By: Top Sellers

Get It Fast

In Stock at Store (31)  
Cuyahoga Falls and nearby stores

Free 2-Day Delivery (114)

Top Filters

Department

Bath

Toilets, Toilet Seats & Bidets

Toilets

Two Piece Toilets

Brand

KOHLER (335)

American Standard (138)

TOTO (99)

Glacier Bay (26)

STERLING (16)

+ See All

Compare

Compare

Compare

Compare

Glacier Bay 2-Piece 1.28 GPF High Efficiency Single Flush Elongated Toilet in White

Model: 1280

1280

\$99.00

Schedule delivery

13 in stock at Cuyahoga Falls

American Standard Champion 4

Max Tall Height 2-Piece High-Efficiency 1.28 GPF Single Flush Elongated Toilet in White with Slow Close Seat

Model: 2556.1280T.020

2556.1280T.020

\$199.00

Schedule delivery

13 in stock at Cuyahoga Falls

Glacier Bay 2-Piece 1.1 GPF/1.6 GPF Complete Elongated Toilet in White, Seat Included

Model: N2316

N2316

\$99.00

Free delivery

28 in stock at Cuyahoga Falls

American Standard Cadet 3 Tall Height Complete 2-Piece 1.28 GPF Single Flush Round Toilet in White with Slow Close Seat

Model: 1277.1280T.020

1277.1280T.020

\$149.00

Schedule delivery

11 in stock at Cuyahoga Falls

Feedback

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

OMNI MARKETING INTERACTIVE

FREE SHIPPING\* + FREE IN STORE PICK UP + FREE RETURNS\*\*

Store Finder | For Pros | Get It Installed | Tool Rental | Credit Center | Savings Center | Project-How-To

More saving. More doing:

My Store Location: South Loop #1950 (Change) Local Ad Help | My Account (Sign in or Register)

SHOP ALL DEPARTMENTS

SEARCH ALL

GO

CART

MY LIST

Home / Bath / Toilets, Toilet Seats & Bidets / Toilets / Two-Piece Toilets

Two-Piece Toilets

Print Page

4 Products Sort By: Top Sellers Results per page: 24

View: Grid List Products: Online In-Store All Products

Select up to 4 items to compare. COMPARE

Select to compare

Select to compare

Select to compare

Select to compare

Pegasus Vicki 2-Piece Round Toilet in White

Pegasus Victoria 2-Piece Round Toilet in White

Pegasus Victoria 2-Piece Round Toilet Beque

Pegasus Vicki 2-Piece Round Toilet in Black

PROJECT IDEAS WITH STYLE

Check out the premier issue of The Home Depot Style Guide - an inspirational, interactive digital magazine full of affordable and achievable home improvement updates.

Look Inside

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

OMNI MARKETING INTERACTIVE

FREE SHIPPING\* + FREE IN STORE PICK UP + FREE RETURNS\*\*

Store Finder | For Pros | Get It Installed | Tool Rental | Credit Center | Savings Center | Project How-To

More saving. More doing.™ My Store Location: **South Loop #1950** (Change) Local Ad Help | My Account (Sign in or Register)

SHOP ALL DEPARTMENTS SEARCH ALL GO CART MY LIST

Home / Bath / Toilets, Toilet Seats & Bidets / Toilets / Two-Piece Toilets





## Two-Piece Toilets

4 Products Sort By: Top Sellers Results per page: 24

View: Grid List Products: Online In-Store All Products

Select up to 4 items to compare. COMPARE

Select to compare Select to compare Select to compare Select to compare

Pegasus Vico 2-Piece Round Toilet in White  
 Pegasus Victoria 2-Piece Round Toilet in White  
 Pegasus Victoria 2-Piece Round Toilet in Black  
 Pegasus Vico 2-Piece Round Toilet in Black

Refine By:  
 Price \$300 - \$400  
 Brand Pegasus  
 Bowl Shape Round  
 Clear All Refinements

PROJECT IDEAS WITH STYLE

Check out the premier issue of The Home Depot Style Guide - an inspirational, interactive digital magazine full of affordable and achievable home improvement updates. Look Inside

Managing Duplicate Content Delivery Copyright 1995-2019. All rights reserved. OMNI MARKETING INTERACTIVE

## Faceted classification leads to duplicate content delivery:

Facet	# of Vocabulary Terms
Type	46
Region	16
Winery	750
Price	6
Ratings	6
Total Terms	824
<b>Total Combinations</b>	<b>19,872,000</b>

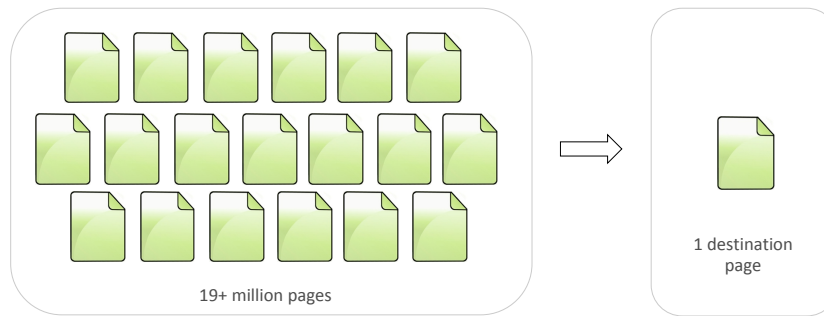
<http://semanticstudios.com/publications/semantics/000003.php>

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Multiple pages → 1 product page:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Is this duplicate content?

BLOGS BY TAG

**CONTENT**

**Oldies But Goodies: Past Content Marketing Trends That Have Staying Power**

High marketing. Searching for that one content marketing trend to...

**Is It Better to be a Salesman or an Educator?**

High marketing. It may have been a common theme to keep trying to sell your products, but you can't...

**How to write relevant, engaging content that gets read**

High marketing. Don't just write for content for the sake of it. Think about your audience and what they are looking for...

BLOGS BY TAG

**CONTENT MANAGEMENT**

**What are Blog Tags, Categories and Subcategories?**

High marketing. Blog tags and categories are a staple in any blog and there's nothing you can do to avoid them...

**How to write relevant, engaging content that gets read**

High marketing. Don't just write for content for the sake of it. Think about your audience and what they are looking for...

**How to write relevant, engaging content that gets read**

High marketing. Don't just write for content for the sake of it. Think about your audience and what they are looking for...

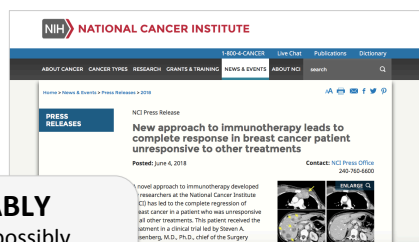
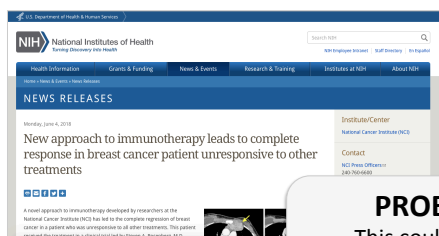
**MAYBE**  
This could possibly be duplicate content to a search engine.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Is this duplicate content?



**PROBABLY**  
This could possibly be duplicate content to a search engine.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Is this duplicate content?



**NO**  
This is unlikely duplicate content to a search engine.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Whenever you hear these words or phrases:

- Personalization
- Site search engine
- Tagging or tagged content
- Faceted classification/navigation
- Sorting options or parameters
- Shared content across network
- Syndication
- Multiple product categories
- Multiple languages
- Aggregation
- https/http

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



Solutions & Best Practices for Managing Duplicate Content Delivery

## WHAT IS DUPLICATE CONTENT?



Copyright 1995-2019. All rights reserved.

**Many people (users) perceive duplicate content...**

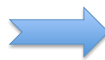


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



**...as an exact match...**



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



**...or a nearly exact match:**



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

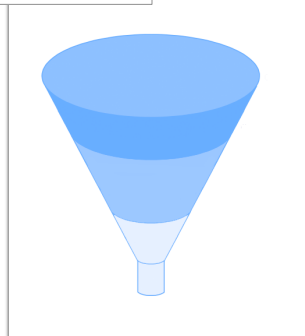


**These perspectives often overlap:**

Human users:



Technology:



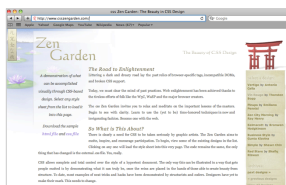
Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Definition is unclear:

- “Exact replicas” is too precise.



- **Source:** Chowdhury, A. et al. (2002). “Collection Statistics for Fast Duplicate Document Detection.” *ACM Transactions on Information Systems*, 20 (2), 171-191.

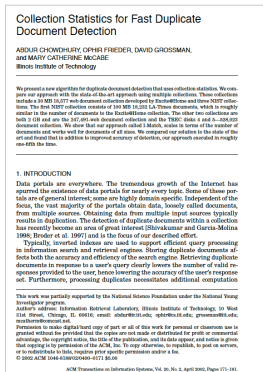
Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## “Resemblance” is more accurate:

- “If a document contains roughly the same semantic content, it is considered a duplicate **whether or not it is an precise syntactic match.**”
- Also available online at:  
<http://ir.cs.georgetown.edu/publications/downloads/p171-chowdhury.pdf>



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Resemblance:

- “The **resemblance** measures whether two (web) documents are roughly the same, that is, they have the same content except for modifications such as formatting, minor corrections, capitalization, web-master signature, logo, etc.”
- Available at:  
<http://cs.brown.edu/courses/cs253/papers/nearduplicate.pdf>
- Google’s label? **Appreciably similar.**

### Identifying and Filtering Near-Duplicate Documents

Andrei Z. Broder\*

AltaVista Company, San Mateo, CA 94402, USA  
andrei.broder@av.com

**Abstract.** The mathematical concept of document resemblance captures well the informal notion of syntactic similarity. The resemblance can be estimated using a fixed-size “sketch” for each document. For a large collection of documents (say hundreds of millions) the size of this sketch is of the order of a few hundred bytes per document. However, for efficient large-scale web indexing it is not necessary to determine the actual resemblance value: it suffices to determine whether newly encountered documents are duplicates or non-duplicates of documents already indexed. In other words, it suffices to determine whether the resemblance is above a certain threshold. In this talk we show how this determination can be made using a “sample” of less than 50 bytes per document.

The basic approach for computing resemblance has two aspects: first, resemblance is expressed as a set (of strings) intersection problem, and second, the relative size of intersections is evaluated by a process of random sampling that can be done independently for each document. The process of estimating the relative size of intersection of sets and the threshold test discussed above can be applied to arbitrary sets, and thus might be of independent interest.

The algorithm for filtering near-duplicate documents discussed here has been successfully implemented and has been used for the last three years in the context of the AltaVista search engine.

#### 1 Introduction

A Communist era joke in Russia goes like this: Leonid Brezhnev (the Party leader) wanted to get rid of the Premier, Aleksey Kosygin. (In fact he did, in

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## According to Google, “copied” content is:

- **Content copied exactly from an identifiable source.** Sometimes an entire page is copied, and sometimes just parts of the page are copied. Sometimes multiple pages are copied and then pasted together into a single page. Text that has been copied exactly is usually the easiest type of copied content to identify.
- **Content that is copied, but changed slightly from the original.** This type of copying makes it difficult to find the exact matching original source. Sometimes just a few words are changed, or whole sentences are changed, or a “find and replace” modification is made, where one word is replaced with another throughout the text. These types of changes are deliberately done to make it difficult to find the original source of the content. We call this kind of content “copied with minimal alteration.”
- **Content copied from a changing source, such as a search results page or news feed.** You often will not be able to find an exact matching original source if it is a copy of “dynamic” content (content that changes frequently). However, we will still consider this to be copied content.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



Solutions & Best Practices for Managing Duplicate Content Delivery

## WHY SHOULD WE CARE?



Copyright 1995-2019. All rights reserved.

### If duplicate content delivery is not managed well:

- **Less qualified site traffic.** Websites often show a dip in site traffic if they deliver duplicate content.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Can experience a dip in site traffic:

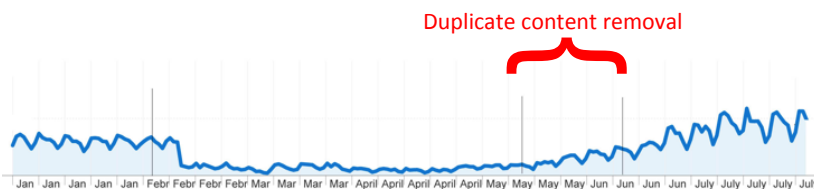


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Or it can be a BIG dip:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



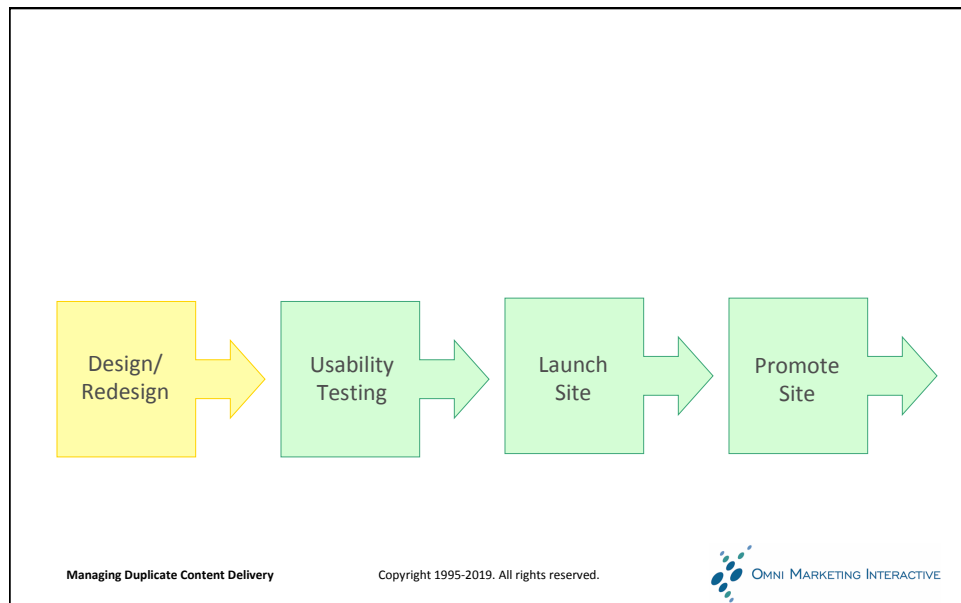
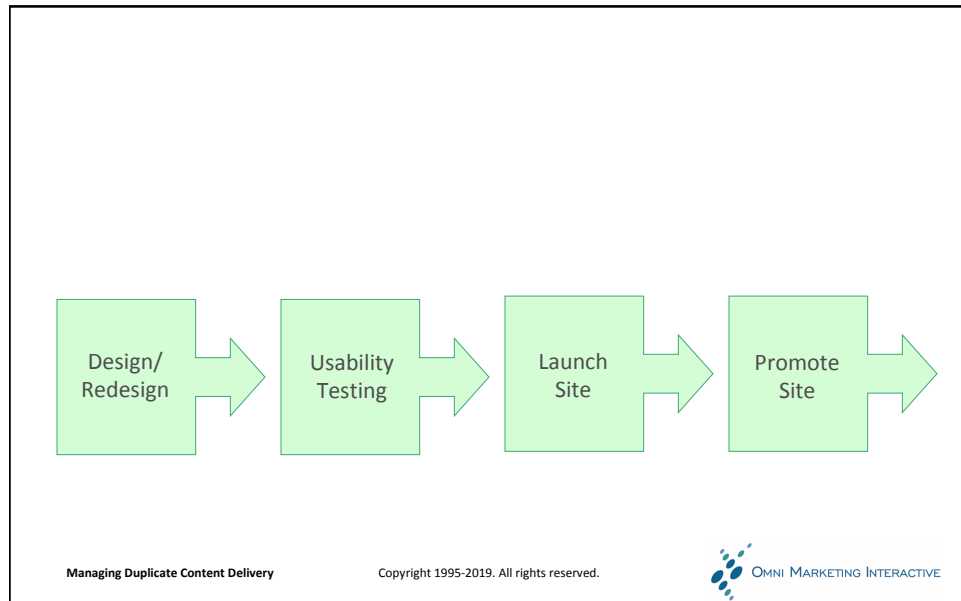
## Important:

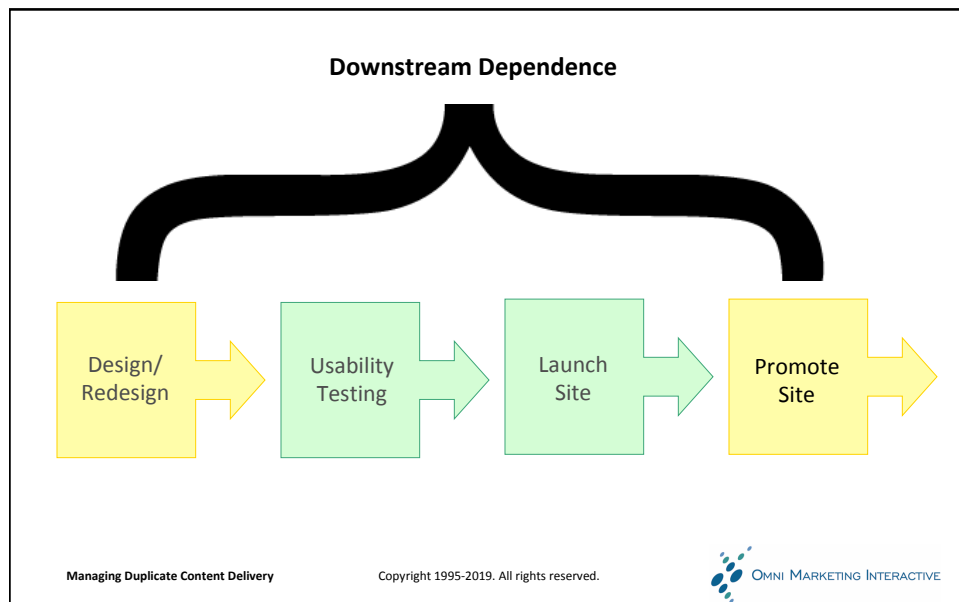


Managing duplicate content delivery is an iterative (ongoing) process, particularly for large sites.

## If duplicate content delivery is not managed well:

- **Less qualified site traffic.** Websites often show a dip in site traffic if they deliver duplicate content.
- **Downstream dependence.** Purchasing a different content management system (CMS) & associated costs; hiring an SEO firm to help manage content delivery.





### If duplicate content delivery is not managed well:

- **Less qualified site traffic.** Websites often show a dip in site traffic if they deliver duplicate content.
- **Downstream dependence.** Purchasing a different content management system (CMS) & associated costs; hiring an SEO firm to help manage content delivery.
- **Low index count.** Less page URLs are available to rank.

### Index count or crawler cap:

- Suppose a website has 1,000 articles:
  - 500 in (X)HTML format
  - Printer-friendly versions of each article on different URLs
- Search engines will most likely filter out the printer-friendly versions of articles from search results.
- Result? Only 500 pages available to rank.

$$\begin{array}{r} 1000 \\ - 500 \\ \hline 500 \end{array}$$

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



### If duplicate content delivery is not managed well:

- **Less qualified site traffic.** Websites often show a dip in site traffic if they deliver duplicate content.
- **Downstream dependence.** Purchasing a different content management system (CMS) & associated costs; hiring an SEO firm to help manage content delivery.
- **Low index count.** Less page URLs are available to rank.
- **Search conversions.** Best converting pages might not appear in search results.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Law site:

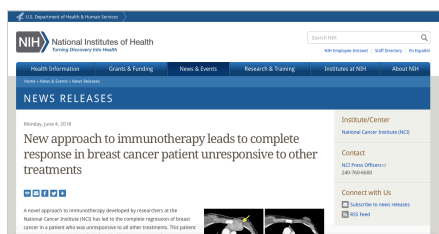
 


- Google autofilled site search engine.
- Law site's search results pages appeared in Google listings.
- Best converting pages? Gone.
- Moral of the story? Be pro-active.

## If duplicate content delivery is not managed well:

- **Less qualified site traffic.** Websites often show a dip in site traffic if they deliver duplicate content.
- **Downstream dependence.** Purchasing a different content management system (CMS) & associated costs; hiring an SEO firm to help manage content delivery.
- **Low index count.** Less page URLs are available to rank.
- **Search conversions.** Best converting pages might not appear in search results.
- **Competition:** Web pages from your shared-content partners' sites (affiliates, syndicates, etc.) have better search engine visibility.

## Which news release has better search engine visibility?



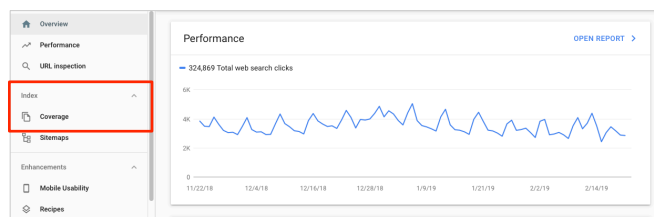
Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## If duplicate content delivery is not managed well:

- **Waste of search engine resources.** Ideally, a search engine wants the original source of content. Less likely to re-crawl website.



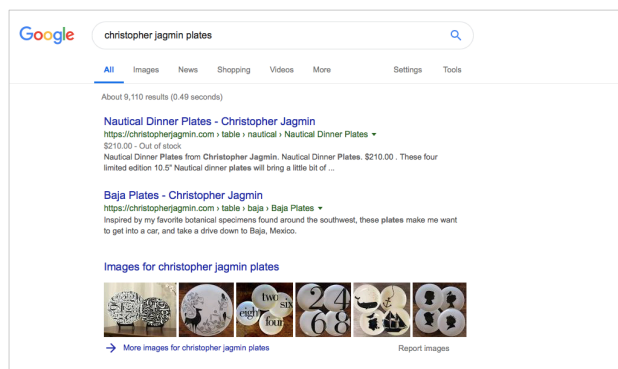
Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## If duplicate content delivery is not managed well:

- **Waste of search engine resources.** Ideally, a search engine wants the original source of content. Less likely to re-crawl website.
- **Poor searcher experience:** Negative brand impact.



The screenshot shows a Google search for "christopher jagmin plates". The search results include several entries for different plate designs, all from the same source (christopherjagmin.com). The results are:

- Nautical Dinner Plates - Christopher Jagmin**: \$210.00 - Out of stock. Nautical Dinner Plates from Christopher Jagmin. Limited edition 10.5" Nautical dinner plates.
- Baja Plates - Christopher Jagmin**: \$210.00 - Out of stock. Baja Plates from Christopher Jagmin. Inspired by my favorite botanical species to get into a car, and take a drive down...
- Even Number Dinner Plates - Christopher Jagmin**: \$130.00 - In stock. Even Number Dinner Plates from Christopher Jagmin. For those mathematicians in the house, these smart plates make a huge, bold statement on your table.
- Little Nautical Plates - Christopher Jagmin**: \$110.00 - In stock. Little Nautical Plates from Christopher Jagmin. Little Nautical Plates. \$110.00. These four limited edition 7.5" Little Nautical plates will bring a little bit of the sea to...
- Products Archive - Christopher Jagmin**: https://christopherjagmin.com/shop/. Products 1 - 12 of 32 - in the Category: Little Whale Plate from Christopher Jagmin ... Little Ship Plate, \$29.00. Little Captain Plate by Christopher Jagmin ...
- Little Ship Plate - Christopher Jagmin**: https://christopherjagmin.com/table/nautical/Little Ship Plate. This limited edition 7.5" little ship plate will bring a little bit of the sea to your dinner ... Little Ship Plate 1 plate 7.5" porcelain. Dishwasher and microwave safe.
- plates Archives - Christopher Jagmin**: https://christopherjagmin.com/table/plates/. Janine Face Plate, \$45.00. 1: 2: 3: ... All work is copyrighted ©Christopher Jagmin 2018. All rights to any art is strictly prohibited, unless there is requested and ...

Managing Duplicate Content Delivery Copyright 1995-2019. All rights reserved. OMNI MARKETING INTERACTIVE

## Worst-case scenario:

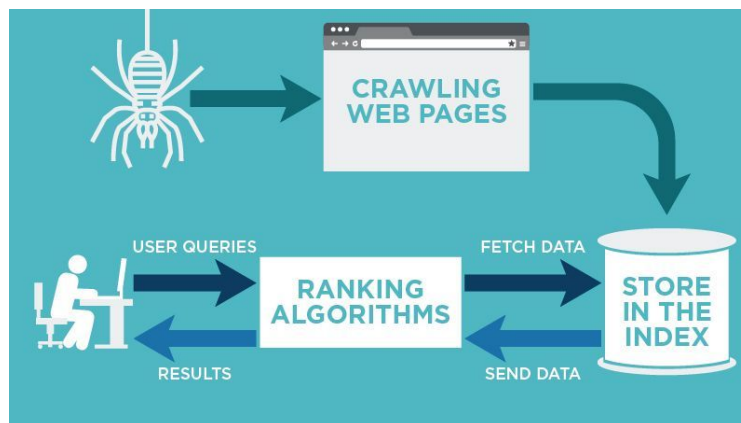
- If search engines determine that duplicate content delivery is deliberate (that is, in order to achieve top search-results positions), **some or all websites can be eliminated from search engines**, especially if Google's Quality Team determines that all of the websites are controlled by the same company.

Solutions & Best Practices for Managing Duplicate Content Delivery

## HOW TO DETERMINE DUPLICATE CONTENT



Copyright 1995-2019. All rights reserved.

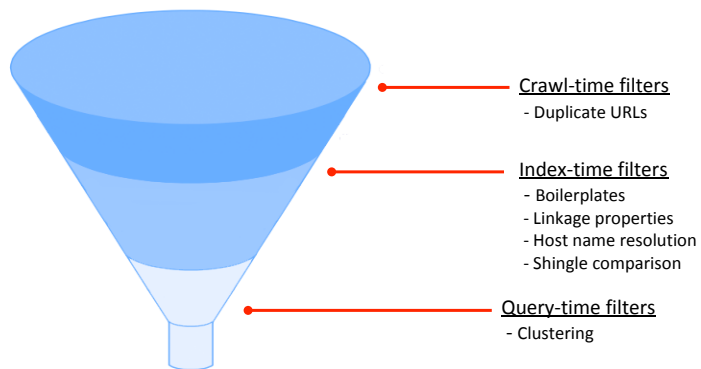


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## How search engines view duplicate content:

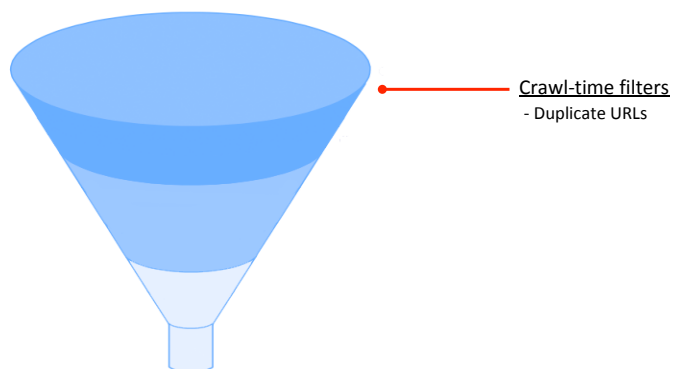


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Crawl-time filters:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Crawl-time filter example

- The following URLs deliver the same content:
  - www.domain.com
  - domain.com/
  - www.domain.com/index.html
- Google (and other search engines) **only want to display one of these URLs.**
- Google (and other search engines) only want to display one of these URLs. Which one is the best choice?
- Best solution?** Consistently link to the same URL on your website. Canonicalize as a supplement.
- If you don't? Google will choose the "best" URL, and it can have a negative impact on web-search rankings.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## This data indicates a huge problem:

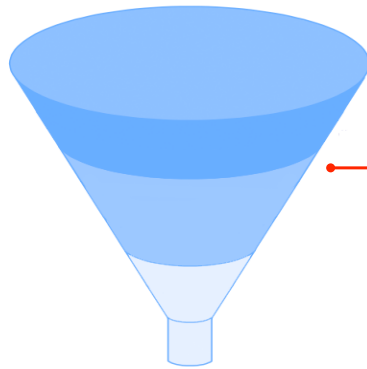


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Index-time filters:



### Index-time filters

- Boilerplates
- Linkage properties
- Host name resolution
- Shingle comparison

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Some index-time filters:

- Boilerplate stripping
- Linkage properties
- Host name resolution
- Shingle & comparison

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

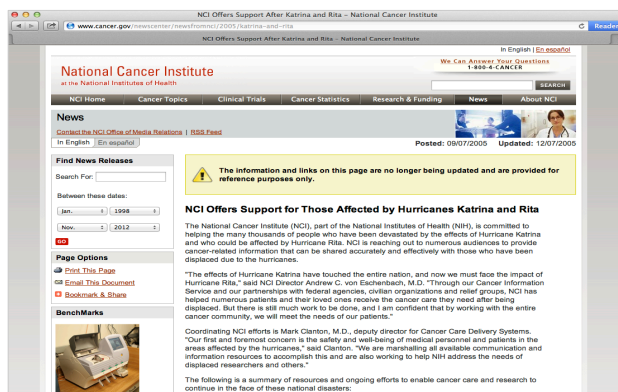


## Boilerplate (template) stripping:

- A **boilerplate** is a section of (X)HTML code that is common to many different documents:
  - Masthead
  - Global navigation
  - Footer
- Search engines remove boilerplate elements to determine a web page's **unique content fingerprint**.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



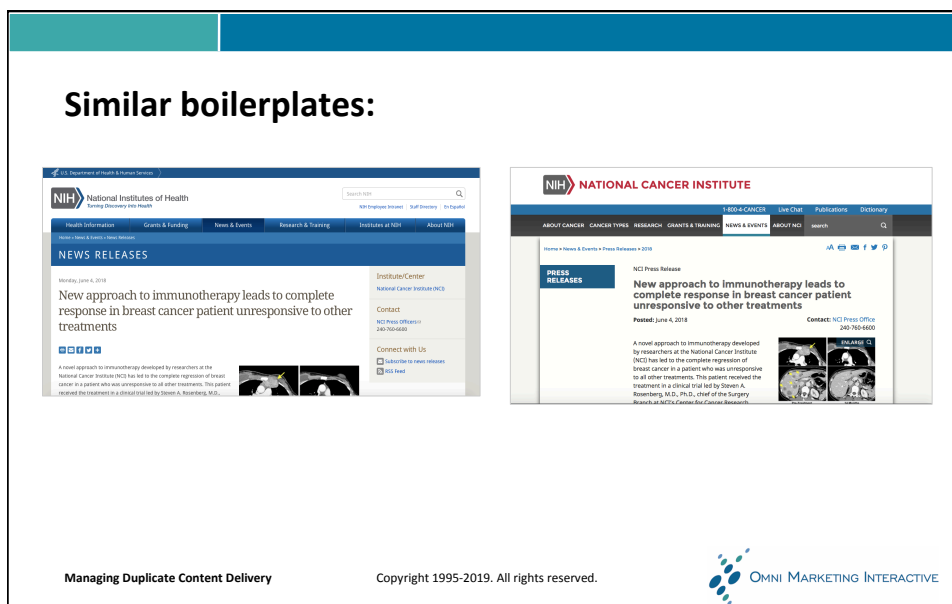
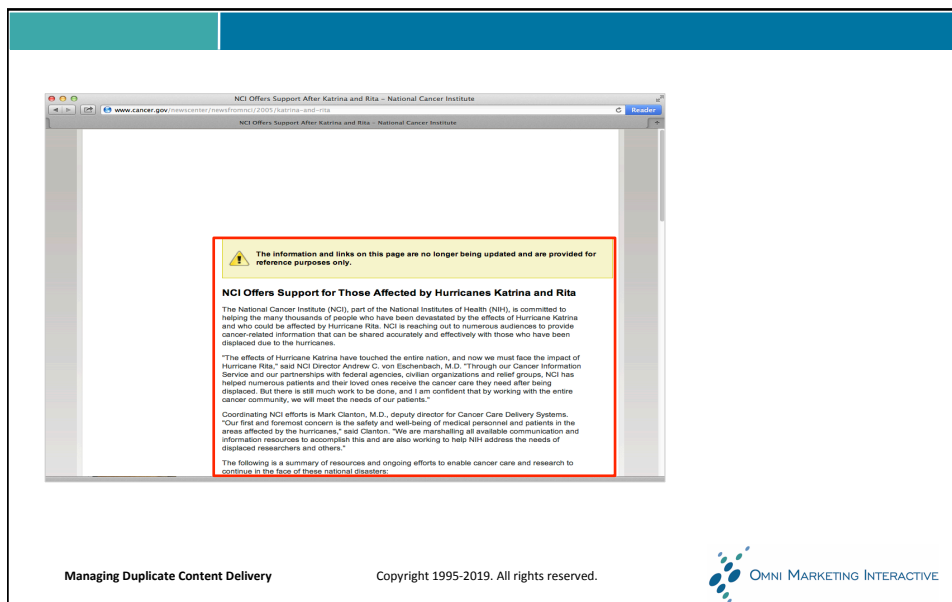
(Also the global footer)



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.





## Linkage properties:

- Inbound links - # of hyperlinks referring to a web page
- Outbound links - # of hyperlinks embedded in a web page

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Different linkage properties:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Host name resolution:

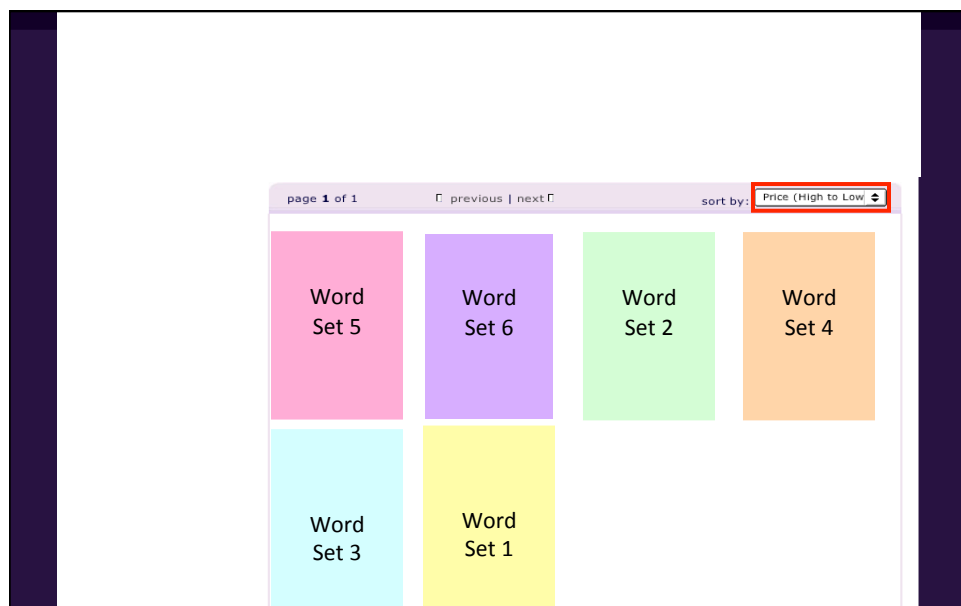
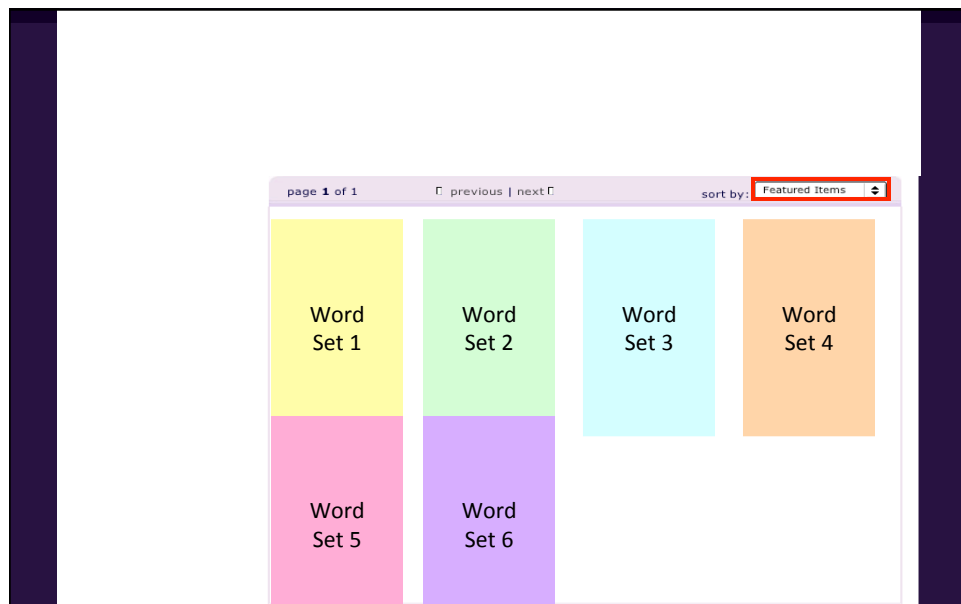
- The host name is the unique name of a machine, such as a web server.
  - Domain name: bmw.com
  - IP address: 160.46.226.165
  - Host name: rrccs-70-63-21-226.central.biz.rr.com
- Also look at other domain/server properties
  - A "C" Block address is based on your IP
- In non-techie language? Search engines want to know who has control over the website(s).

## Shingle comparison:

- Every web document has a unique content signature or "fingerprint".
- Content is broken down into sets of word patterns.
- Groups of adjacent words, called shingles, are compared for similarity (shingle footprint).
- **More shingles = more similarity.**
  - » For example: Once upon a midnight dreary, while I pondered weak and weary.

Once upon a  
 upon a midnight  
 a midnight dreary  
 midnight dreary while  
 dreary while I  
 while I pondered  
 I pondered weak

- Position of text on page has little weight.





### Both pages are similar:

- 2 web pages with unique URLs (web addresses).
- The same word sets are available on both web pages.
- Is this duplicate content?

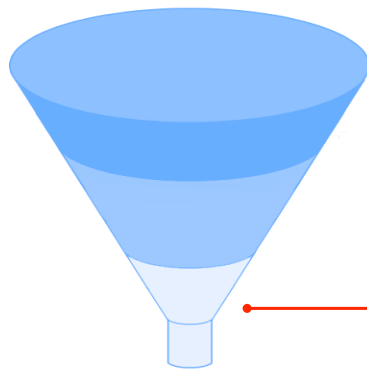
### In other words:

Sort By: Most Popular ▾

 LIST  DETAIL  SCREEN

View:  Grid |  List Sort By: Top Sellers ▾ Results Per Page: 24 ▾

## Query-time filters:



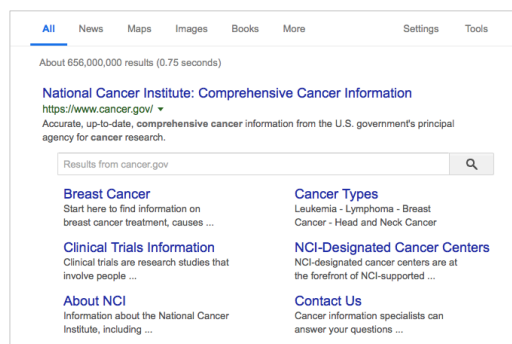
Query-time filters  
- Clustering

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Query-time filter example:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Quick duplicate content checklist:

Questions to Ask:	Yes	No
<b>Boilerplate (templates)</b> Are the page boilerplates the same or very similar?	✓	
<b>Linkage properties (inbound and outbound links)</b> Are the linkage properties the same or very similar?	✓	
<b>Host name resolution</b> Does the same company/organization control the content of both pages?	✓	
<b>Shingles/super-shingles/mega-shingles (word sets)</b> Are the same shingles used on both pages? (Note: arrangement of shingles does not matter – re-sorting the same word sets is still considered duplicate content.)	✓	

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



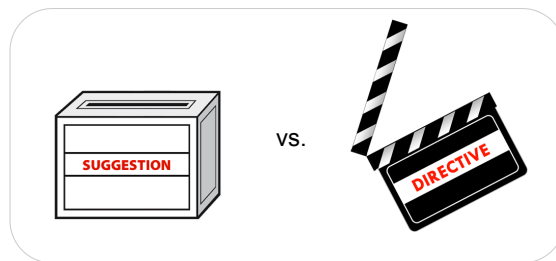
Solutions &amp; Best Practices for Managing Duplicate Content Delivery

## DUPLICATE CONTENT SOLUTIONS



Copyright 1995-2019. All rights reserved.

## Before we go over 8 items:



Signal = maybe

Directive = yes

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Duplicate content solutions:

1. Information architecture (IA) & site navigation
2. Robots.txt
3. Robots meta tag
4. Canonicalization
5. Redirects (301)
6. NOFOLLOW attribute
7. Web search engine webmaster tools
8. Sitemap (XML)

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Site navigation & IA:

- Are you linking consistently to the same URLs throughout the site with unique anchor text?
- Remember, search engines want the shortest URL that delivers unique content.
  - Communicate aboutness
  - Communicate strong information scent
  - Does not have to reflect taxonomy
  - Minimize use of the following characters: &, ?, =, \$, +, %
- Example (personalized content):
  - <http://www.domain.com/cat/product.php?sessionId=CVX3WR>

**DeepCrawl** @DeepCrawl · 12 Nov 2018  
 .@JohnMu says when using subfolders for international **site structure**, have country before language. 🌐

Want to know why? Our #DeeplyNotes recap of the latest Webmaster Hangouts has all the details.

**Google Webmaster Central Hangout 30th October 2018**



**Google Webmaster Hangout Notes: October 30th 2018 - DeepCrawl**  
 Notes from the Google Webmaster Hangout with John Mueller on the 30th of October 2018.  
[deepcrawl.com](https://www.deepcrawl.com)

<https://www.deepcrawl.com/blog/news/google-webmaster-hangout-notes-october-30th-2018/>



John 🐦 @JohnMu · 29 Jul 2018

Replying to @up\_ka\_shora

It helps Google if there's a clear **structure** on a page, with headings & content, but we're not going to count it against a **site** if they improvise / get it wrong. That said, a clear semantic **structure** of a page can also make sense outside of search.

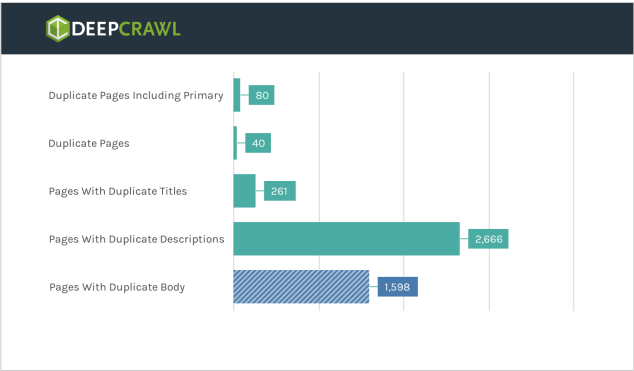
2 14 26

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

 OMNI MARKETING INTERACTIVE


## Also look at your page content:



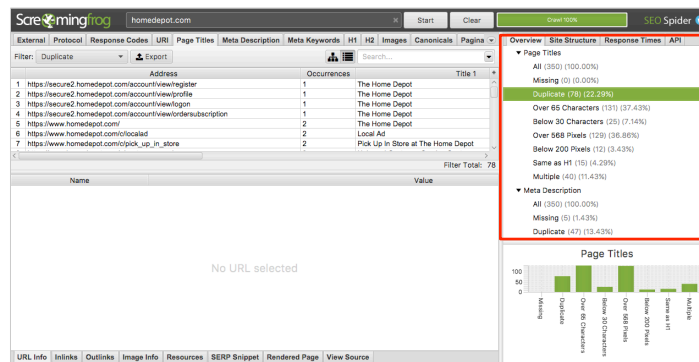
Category	Count
Duplicate Pages Including Primary	80
Duplicate Pages	40
Pages With Duplicate Titles	261
Pages With Duplicate Descriptions	2,666
Pages With Duplicate Body	1,598

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

 OMNI MARKETING INTERACTIVE

## Expect some duplication in paginated content:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Metadata is usually more important with a database or knowledgebase search:

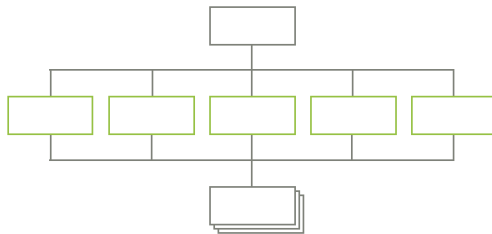


Figure 16 – 6. A database pattern

Image from: Spencer, D. (2010). A Practical Guide to Information Architecture (Vol. 1). Penarth: Five Simple Steps.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Robots.txt:

- Are you preventing the page from being spidered?
- Recommended for site search results.

User-agent: \*  
Disallow: /search/

- Use wildcard, if necessary:

Pattern matching

\* = matches any sequence of characters

\$ = matches the end of a URL

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



**Screaming Frog SEO Spider** - homedepot.com

Filter: All | Export | Search...

Address	Content	Status Code
1 http://homedepot.com/	text/html	301 Moved
2 https://www.homedepot.com/	text/html; charset=UTF-8	200 OK
3 https://www.homedepot.com/localhost	text/html; charset=UTF-8	200 OK
4 https://www.homedepot.com/click_up_in_store	text/html; charset=UTF-8	200 OK
5 https://www.homedepot.com/content/css/desktop-main-no-tilapp.css?v=0...	text/css; charset=UTF-8	200 OK
6 https://www.homedepot.com/services/	text/html; charset=utf-8	200 OK
7 https://www.homedepot.com/_bml/cbdt-1-35	application/javascript	200 OK

Filter Total: 399

No URL selected

URL Info | Inlinks | Outlinks | Image Info | Resources | SERP Snippet | Rendered Page | View Source

**Summary**

- Total URI Encountered: 500
- Total Internal Blocked by robots.txt: 1
- Total External Blocked by robots.txt: 0
- Total URI Crawled: 499
- Total Internal URI: 399
- Total External URI: 101

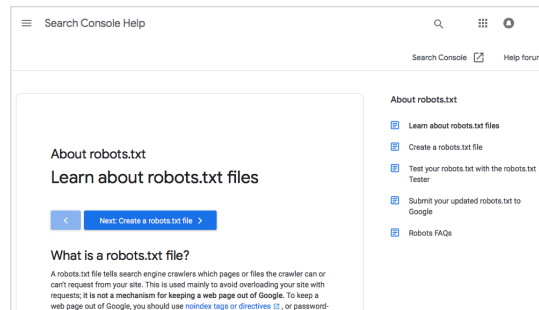
**Internal**

- All (399) (100.00%)
- HTML (343) (85.96%)
- JavaScript (11) (2.76%)
- CSS (4) (1.00%)
- Images (16) (4.01%)
- Other
- Unknown

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

## To Google, it's a signal, not a directive:



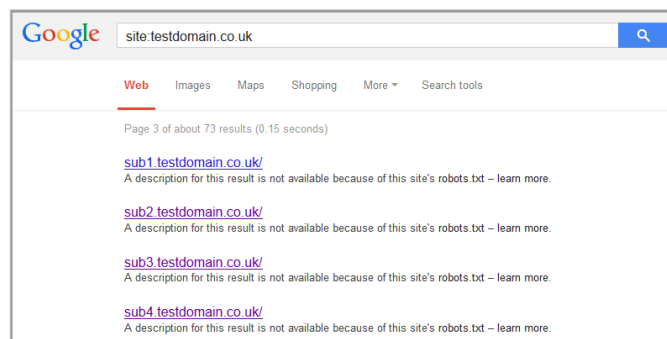
<https://support.google.com/webmasters/answer/6062608>

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## If anyone links to content:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Robots meta tag:

- If articles are shared across your network of sites, are you implementing NOINDEX, NOFOLLOW appropriately?

```
<META NAME="ROBOTS" CONTENT="NOINDEX, FOLLOW" />
```

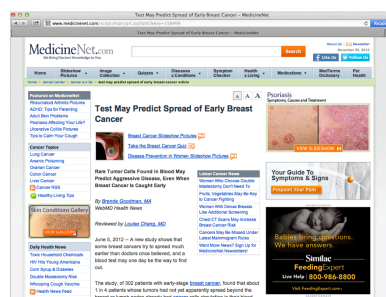


Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## For articles across network:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Canonicalization:

- These URLs are all different to a search engine:

- www.example.com
- www.example.com/
- example.com
- example.com/
- www.example.com/index.html
- www.example.com/index.aspx
- example.com/index.html
- example.com/default.cfm

**Settings**

Geographic target ☐ Target users in: United States


Preferred domain

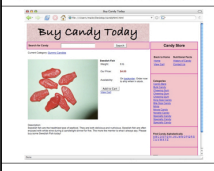
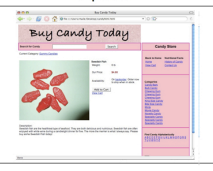
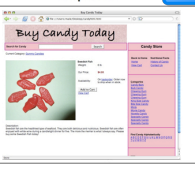
- ☐ Don't set a preferred domain
- ☒ Display URLs as www.searchenginebook.com
- ☐ Display URLs as searchenginebook.com

- In the <head> portion of a page:

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

## Having control over URL structure is extremely important:



Webmaster's intended display URL	Duplicate	Duplicate
		
http://www.example.com/product.php?item=swedish-fish	http://www.example.com/product.php?category=gummy-candy&item=swedish-fish&affiliateid=ABCD	http://www.example.com/product.php?item=swedish-fish&trackingid=1234&sort=price&sessionid=5678

## Not sure of a canonical URL? Step 1:

The screenshot shows the Screaming Frog SEO Spider tool interface. The 'Canonicals' tab is active, displaying a summary of canonical issues. A red box highlights the 'Canonicals' summary section, which includes the following data:

Category	Count	Percentage
All	350	100.00%
Contains Canonical	341	97.43%
Self Referencing	372	89.14%
Canonicalised	290	8.29%
Missing	39	11.14%
Multiple	41	11.71%
Non-Indexable Canonical	0	0.00%

The interface also shows a table of URLs with their canonical status and a 'Filter Total: 350' indicator.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Not sure of a canonical URL? Step 2:

### URL Inspection Tool

#### About the URL Inspection tool

The URL Inspection tool provides information about Google's indexed version of a specific page. Information includes AMP errors, structured data errors, and indexing issues.

#### Features:

- **Inspect an indexed URL:** Retrieve information about Google's indexed version of your page.
- **Inspect a live URL:** Test whether a page on your site is able to be indexed.
- **Request indexing for a URL:** You can request that an inspected URL be crawled by Google.
- **View a rendered version of the page:** See a screenshot of how Googlebot sees the page.
- **View loaded resources list, JavaScript output, and other information:** See a list of resources, page code, and more information by clicking the more information link on the page verdict card.

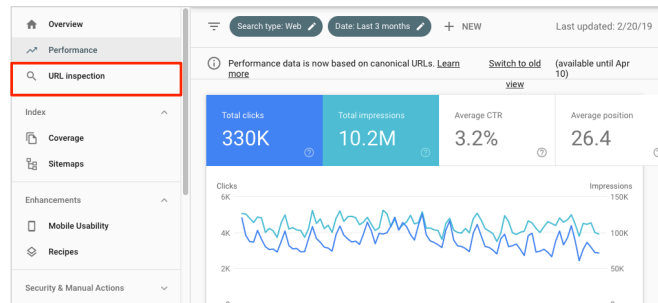
<https://support.google.com/webmasters/answer/9012289>

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Not sure of a canonical URL? Step 3:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



**Google Webmaster Central Blog**  
Official news on crawling and indexing sites for the Google index

### Consolidating your website traffic on canonical URLs

Wednesday, February 06, 2019

In Search Console, the [Performance report](#) currently credits all page metrics to the exact URL that the user is referred to by Google Search. Although this provides very specific data, it makes property management more difficult; for example: if your site has mobile and desktop versions on different properties, you must open multiple properties to see all your Search data for the same piece of content.

To help unify your data, Search Console will soon begin assigning search metrics to the (Google-selected) [canonical URL](#), rather than the URL referred to by Google Search. This change has several benefits:

Hey! Check here if your site is mobile-friendly.

Search blog ...

Labels

Archive

Feed

<https://webmasters.googleblog.com/2019/02/consolidating-your-website-traffic-on.html>

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## 301 redirects:

- If the same URLs are delivering identical content, are you selecting the best URL (usually the shortest one) and 301 redirecting the other URLs to that content?
- Use only one redirect!
  - Old page
  - New/destination page

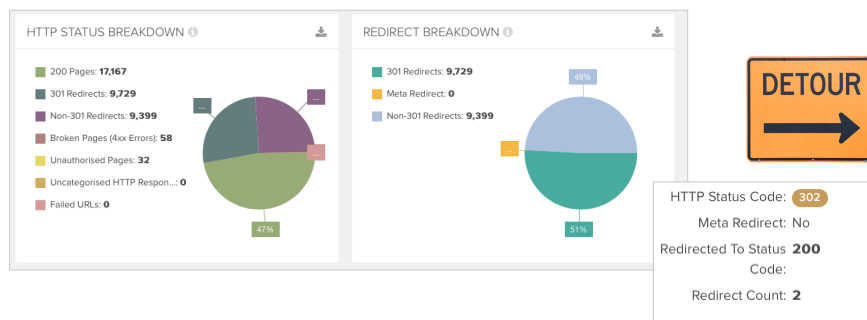
The screenshot shows the USPS 'The Official Change of Address Form'. It includes sections for 'Please enter your name.', 'Enter your old address.', and 'Enter your new address.'. Each section has fields for Street, City, State, and ZIP Code. There is also a 'Your Information' sidebar on the right with fields for Forwarding Date, Permanent or Temporary Move, and Type of Move.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



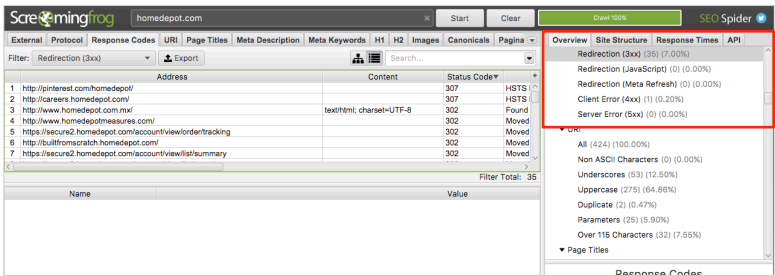
## Non-301 Redirects



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



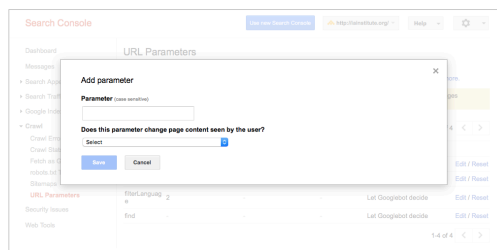


**Managing Duplicate Content Delivery**

Copyright 1995-2019. All rights reserved.

OMNI MARKETING INTERACTIVE

## Web search engine webmaster tools (Google Search Console):



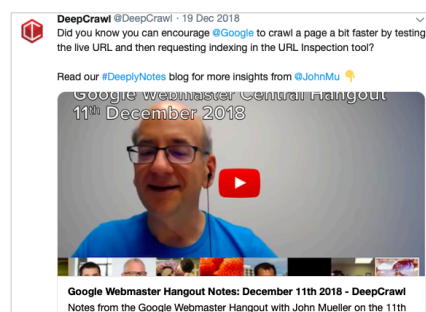
<https://support.google.com/webmasters/answer/6080548>

## NOFOLLOW attribute:

- Are you using the NOFOLLOW attribute on anchor links that you cannot vouch for (such as blog comments or reviews)?
- Are you using the NOFOLLOW attribute on forms that search engines might "accidentally" submit?
  - `<a href="send-email.php" rel="nofollow">Send E-mail</a>`
- Paid links.
- Signal, not a directive.

## XML sitemap:

- Not a wayfinder site map.
- An XML sitemap is not a REQUIREMENT on any website. It is recommended for sites whose content changes quickly. The site might get crawled faster.
- <https://www.deepcrawl.com/blog/news/google-webmaster-hangout-notes-december-11th-2018/>



## Important!



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Be consistent!

- For example, do not robots exclude (either via robots.txt or the robots meta tag) a URL and then include the URL in the XML sitemap.
- Do not implement a NOFOLLOW attribute on important pages. For example, a Privacy Policy might not be important from a ranking perspective, but it is important to site visitors.
- Lack of consistency? Undesirable results.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## To remember:

- **Be proactive** – Come up with a clear URL (web address) structure & information architecture so that it is easier to not deliver duplicate content.
- Duplicate content management is an ongoing process.
- Use web analytics software & other tools to determine the pages that have the highest conversion rates. Exclude the duplicates from crawling.
- Do not use robots exclusion, redirects, NOFOLLOW attribute, etc. as a substitute for a poor or substandard information architecture – degrades the searcher experience.
- **Be consistent** – If you consistently provide search engines with clues as to which pages you want to appear in search results, then they are more likely to select the “right” page.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



Solutions & Best Practices for Managing Duplicate Content Delivery

## FACETED CLASSIFICATION



Copyright 1995-2019. All rights reserved.

## What I do for faceted classification:

### 1. Start with the 3 primary ways users organize & label content.

- Open card-sort test (best)
- Closed card-sort test or tree test
- First click test

### 2. Leave site alone. Learn how web search engines crawl & index site.

- If you do this, do NOT submit an XML sitemap.
- A wayfinder site map & site index are perfectly acceptable.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



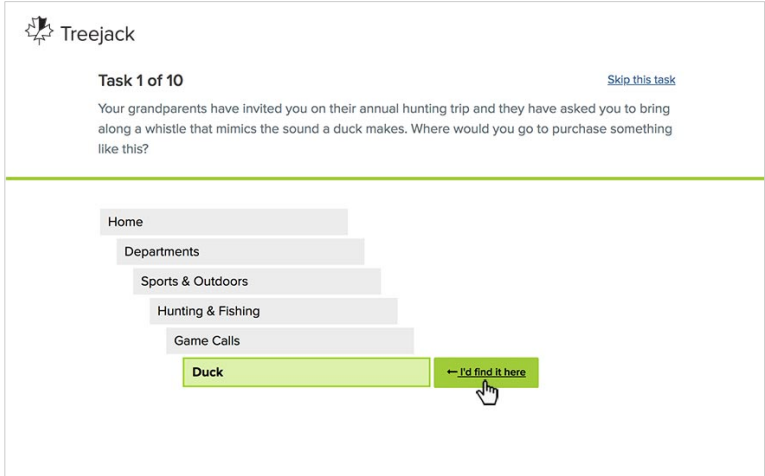
Will Yelp remove reviews that contain offensive content?  Can I report a photo or video that violates my privacy rights?  How can my business be featured in an upcoming edition of the Weekly Yelp?  Should I report a business that is trying to pay people to write positive reviews?  What's the deal with those companies that claim to be able to help me manage my reputation on Yelp?  How do I post a Yelp deal or	<b>Business account owners</b>  How do I post my business's menu to Yelp?  How can I change which photos of my business appear first?  How do I claim a business page that has already been claimed?  Can I remove my business page from Yelp?	<b>Deals and vouchers</b>  How can I reach out to customers who've bought my Yelp deal?  How do I remove a Yelp deal?	<b>Reviews and</b>  Can I use reviews or my average star rating on Yelp in marketing materials for my business?  How do I respond to reviews of my business?  Do reviews that aren't currently recommended impact my business's star rating?  Can I sue Yelp for a bad review?
---	--	---	--

<https://www.optimalworkshop.com/101/card-sorting>

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.





**Task 1 of 10** [Skip this task](#)

Your grandparents have invited you on their annual hunting trip and they have asked you to bring along a whistle that mimics the sound a duck makes. Where would you go to purchase something like this?

Home  
Departments  
Sports & Outdoors  
Hunting & Fishing  
Game Calls  
**Duck** ← I'd find it here

Managing Duplicate Content Delivery Copyright 1995-2019. All rights reserved. OMNI MARKETING INTERACTIVE

## What I do for faceted classification (cont'd):

### 3. Continue to leave site alone.

- If your site search results are accurate? Good.
- If your web search results are accurate? Good. If not? Go to Step 4.

### 4. Find where duplicate content exists that might be causing problems.

- Web analytics software
- Webmaster tools at web search engines
- Paid search analytics software

## What I do for faceted classification (cont'd):

### 5. Fix duplicate content issues.

- Be consistent!
- You might want to hire an objective consultant to double check that there is consistency.

### 6. Lather, rinse, repeat.

Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.

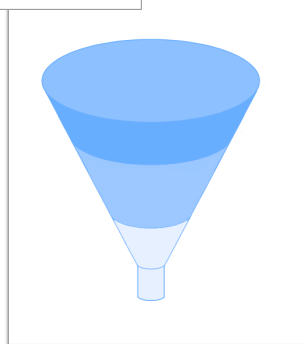


## And remember:

Human users:



Technology:



Managing Duplicate Content Delivery

Copyright 1995-2019. All rights reserved.



## Questions?



Shari Thurow, Founder & SEO Director  
Omni Marketing Interactive

[sthurow@search-usability.com](mailto:sthurow@search-usability.com)



@sharithurow