Solutions & Best Practices for Managing Duplicate Content Delivery

By
Shari Thurow, Founder & SEO Director

Workshop agenda:

• Identifying duplicate content
• What is duplicate content?
• Why should we care?
• How to determine duplicate content
• Duplicate content solutions
• Questions & answers
Main goal of this workshop – mental model:

Human users:

Technology:

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.
Solutions & Best Practices for Managing Duplicate Content Delivery

IDENTIFYING DUPLICATE CONTENT

Search engines do not determine duplicate content by appearance.
Is this duplicate content?

YEAS
This is duplicate content to a search engine.
Faceted classification leads to duplicate content delivery:

<table>
<thead>
<tr>
<th>Facet</th>
<th># of Vocabulary Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>46</td>
</tr>
<tr>
<td>Region</td>
<td>16</td>
</tr>
<tr>
<td>Winery</td>
<td>750</td>
</tr>
<tr>
<td>Price</td>
<td>6</td>
</tr>
<tr>
<td>Ratings</td>
<td>6</td>
</tr>
<tr>
<td>Total Terms</td>
<td>824</td>
</tr>
<tr>
<td>Total Combinations</td>
<td>19,872,000</td>
</tr>
</tbody>
</table>

Is this duplicate content?

MAYBE
This could possibly be duplicate content to a search engine.
Is this duplicate content?

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

NO
This is unlikely duplicate content to a search engine.
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

Solutions & Best Practices for Managing Duplicate Content Delivery

WHAT IS DUPLICATE CONTENT?
Many people (users) perceive duplicate content...

...as an exact match...
...or a nearly exact match:

These perspectives often overlap:

Human users:

Technology:
Definition is unclear:

- “Exact replicas” is too precise.


“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.”

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](http://cs.brown.edu/courses/cs253/papers/nearduplicate.pdf)

- Google’s label? **Appreciably similar.**

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.
Solutions & Best Practices for Managing Duplicate Content Delivery

WHY SHOULD WE CARE?

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.
Can experience a dip in site traffic:

Or it can be a BIG dip:

Duplicate content removal
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.
- **Lowers 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.

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.

- **Lowers index count.** Less page URLs are available to rank.

- **Search conversions.** Best converting pages might not appear in search results.
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.
• **Lowers 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?

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.
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.
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
How search engines view duplicate content:

Crawl-time filters:
- Duplicate URLs

Index-time filters:
- Boilerplates
- Linkage properties
- Host name resolution
- Shingle comparison

Query-time filters:
- Clustering
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.

This data indicates a huge problem:

![Universal Crawl Page Gaps](image)

![Universal Crawl URL Breakdown](image)
Index-time filters:

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

Some index-time filters:

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

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

- 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:
Query-time filters:

Query-time filter example:

Query-time filters
- Clustering

National Cancer Institute: Comprehensive Cancer Information
https://www.cancer.gov

Breast Cancer
Cancer Types
Clinical Trials Information
About NCI
Contact Us

Results from cancer.gov

About 656,036,918 results (0.75 seconds)
Quick duplicate content checklist:

<table>
<thead>
<tr>
<th>Questions to Ask</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boilerplate (templates)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Are the page boilerplates the same or very similar?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linkage properties (inbound and outbound links)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Are the linkage properties the same or very similar?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Host name resolution</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Does the same company/organization control the content of both pages?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shingles/super-shingles/mega-shingles (word sets)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>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.)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Before we go over 8 items:

Signal = maybe  
Directive = yes

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)
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):
Also look at your page content:

### Duplicate Pages Including Primary
- 250

### Duplicate Pages
- 100

### Pages With Duplicate Titles
- 200

### Pages With Duplicate Descriptions
- 2266

### Pages With Duplicate Body
- 1500
Expect some duplication in paginated content:

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

Robots.txt:

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

```plaintext
User-agent: *
Disallow: /search/
```

- Use wildcard, if necessary:

  Pattern matching
  
  * = matches any sequence of characters
  $ = matches the end of a URL
To Google, it’s a signal, not a directive:

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

If anyone links to content:
Robots meta tag:

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

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

For articles across network:
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

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

Managing Duplicate Content Delivery

Not sure of a canonical URL? Step 2:

Managing Duplicate Content Delivery

https://support.google.com/webmasters/answer/9012289
Not sure of a canonical URL? Step 3:

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

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.
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.
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.
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.

And remember:

Human users:

Technology:
Questions?

Shari Thurow, Founder & SEO Director
Omni Marketing Interactive
sthurow@search-usability.com

@sharithurow