Advanced Googling*

David Durden | Data Services Librarian, University of Maryland

February 14, 2018 | LBSC 702: User Instruction

> *...you should please only use "Google" when you're actually referring to Google Inc. and our services.*[1]

1. https://googleblog.blogspot.com/2006/10/do-you-google.html

Topics

- <meta/>, microdata, and how web indexing works
- Think like an algorithm: Search engine optimization
- Google query syntax
- Extending Google search functionality

Web Indexing: How does it work?

- Search engines use web crawlers
- They are seeded with a list of URLs
- URLs are parsed for metadata/key terms
- Data are stored, indexed, and ranked

Web Indexing By the Numbers

- The indexable web consists of ~11.5 billion pages
- Across different SE's, 40-70% of the web is indexed
- Google claims to index ~8 billion pages

That's, like, so <meta/>

Well constructed HTML will have meta-tags that contain page-level descriptive info

- Page title
- Page description
- Creator supplied keywords
- Technical details

Microdata: Attributes for HTML

- HTML tags are mostly structural; Microdata adds descriptive context to HTML
- Schema.org provides authority control for strucured data vocabularies on the web
- One example of linked data implemented successfully

Original HTML:

```
<div>
 <h1>Avatar</h1>
 <span>Director: James Cameron (born August 16, 1954)</span>
 <span>Science fiction</span>
 <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

With Microdata:

```
<div itemscope itemtype ="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <span>Director: <span itemprop="director">James Cameron</span>
        (born August 16, 1954)</span>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer
  </a>
</div>
```

2. "Getting started with schema.org using Microdata." http://schema.org/docs/gs.html

Microdata Powers the Knowledge Graph

- Authoritative* websites contribute data to Google's Knowledge Graph Database

3. *Authoritative in this case means that a website is professionally operated, dedicated to producing quality content, and has a history of reliability/verifiability. Examples include dictionary.com, wikipedia.com, *.gov, etc.

Search Engine Optimization: Gaming the System

> "*The process of maximizing the number of visitors to a particular website by ensuring that the site appears high on the list of results returned by a search engine.*"[4]

4. Google search for "search engine optimization definition"; definition taken from Google Knowledge Graph.

SEO is Partially Responsible for Internet Word Salad[5]

- Titles as keyword lists
- Text is more "machine readable" than human readable
- This practice stems from user-generated content where the user does not have access to HTML or metadata (e.g., social media, YouTube, etc.)

5. For a disturbing view of SEO and algorithmically generated content, read James Bridle's "Something's Wrong on the Internet" https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

Cars 2 Silver Lightning McQueen Racer Surprise Eggs Disney Pixar Zaini Silver Racers by ToyCollector

29,765,731 views

LIKE DISLIKE SHARE

Up next AUTOPLAY

Cars 2 Carrera Go! Slot Racing Track Silver Lightning McQueen
Blu Toys Club Surprise
1.4M views

Disney Pixar Cars3 Toy Movie Big Mack Truck Gale Beaufort
おもちゃんねるん Omotyanner
21M views

Cars carry case car toy & 48 My Collection TOMICA Cars
mania japansong
4.7M views

Disney Cars Toys GIANT EGG SURPRISE OPENING Lightning
Awesome Toys Collectors (Gia
Recommended for you

100+ cars toys GIANT EGG SURPRISE OPENING Disney
Ryan ToysReview
839M views

Google Search Operators

Boolean Concepts

- Values: *true/false*
- Operators: *OR, AND, NOT*
- Expressions: *pet NOT cat*

Google Booleans

- AND is implied: *maryland election 2016 = maryland AND election AND 2016*
- OR is explicit: *walking OR running*
- NOT is expressed using 'minus': *vans -shoes*

Search Operators

Operator	Definition	Usage
..	Range	9.99..20
*	Wildcard	2 * to midnight
@	Search Social Media	@twitter dogs
#	Search Hashtags	#funny

More Operators

Operator	Definition	Usage
( )	Order of Evaluation	(dogs OR cats)
in	Conversions	100 mph in kph
$	Search Prices	pants $20
" "	Exact Match	"david durden umd"

Even More Operators

Operator	Definition	Usage
site:	Limits search by domain type	site:lib.umd.edu site:.edu site:uk
inurl:	Searches URL strings	inurl:2015
filetype:	Limits by file type	the republic filetype:pdf
intext:	Searches webpage for any instance of text	intext:data management
define:	Defines query string	define:query

Smooth Operators

Operator	Definition	Usage
related:	Finds similar websites	related:nytimes.com
cache:	Returns cached copy of page	cache:domain query
info:	Returns info about a domain	info:lib.umd.edu
safesearch:	Enables safe search	safesearch:query

Query Construction Strategies

Think Compartmentally

- Identify and augment
- Refine and iterate

Example:

1. taco recipe

2. taco recipe site:mx[6]

3. taco recipe site:mx -site:tablespoon.com.mx

6. ISO 3166-1 alpha-2 https://en.wikipedia.org/wiki/ISO_3166-1

Examine URL Structure

Identify elements from a site's URL that you want filtered out of or included in results

```
https://www.washingtonpost.com/business/economy/white-house-budget-proposes-increase-to-defense-spending-and-cuts-to-safety-net-but-federal-deficit-would-remain/2018/02/12/f2eb00e6-100e-11e8-8ea1-c1d91fcec3fe_story.html?hpid=hp_rhp-top-table-main_budget-1142am%3Ahomepage%2Fstory&utm_term=.7de4b1e80578
```

Example:

budget site:washingtonpost.com -inurl:business -inurl:economy

Tips for Using Google More Reliably

- Use an incognito or private browser
- Adjust search settings
- Know when to switch to Google Scholar
- Know when to use the Advanced Search form
- Setup your digital workspace

Handy Chrome Extensions for Research

- Zotero Connector
- Save to Google Drive
- Google Scholar Button
- Evernote Web Clipper

The Power of the Reverse Image Search

Drop any image into the search bar of images.google.com to lookup that image

Google URLs

Base	Advanced
https://google.com	https://google.com/advanced_search
https://images.google.com	https://google.com/advanced_image_search
https://scholar.google.com	*Menu option only*
https://patents.google.com	https://patents.google.com/advanced

💾 Thanks!

💻 https://lib.umd.edu/data

✉️ lib-research-data@umd.edu

✉️ durden@umd.edu