Chapter 2

Exposing Content on the Web

To understand how anything is exposed on the web, it's useful to understand how Google, the most widely used of the search engines, indexes and ranks the content that it gathers.

Google wasn't the first company on the web to provide search results across Internet content, but soon after its late 1990s debut, its search product became almost exclusively associated with searching and finding exactly the thing that the searcher was looking for. It beat competitors with colorful names like AltaVista, Yahoo!, and HotBot by providing the search tool that generally found the right thing, ranking that right thing at the top of the list of results, and doing it all in less than a second. The market quickly perceived that Google Search did that better than anybody, and Google has generally retained that position today. It has even influenced spoken languages as its success inspired the new verb *to google* as a standard way of indicating "to search for answers on the web."

Google Search Methodology

Google's science of crawling, indexing, and ranking webpages is well understood insofar as Google explains the mechanics to specialists and the general public. What follows below is a high-level view of how those mechanics work based on the information that Google makes public—it is necessarily simplified to provide the basics and to illustrate that how to make content visible on the web is a known art and science.

To encourage the creation of crawlable and indexable websites, Google provides good information and even technical tools so individuals and companies can design their webpages for the best possible results. It is in Google's interest to encourage good behavior in website design so it can maximize the quality of its search results. Google also wants to minimize the amount of work it has to do to prevent bad behavior. The behavior it wants to discourage is where web designers try to game the PageRank system to advantage their own content. Google discourages that behavior through sophisticated algorithms and punitive removal of content from search results. Those removed from results have to petition Google to have their content displayed again, and Google's engineers have to be convinced that the behavior was not intentional and corrected before they are cleared.

In addition to crawling the open web, Google will seek out partners and create formal partner contracts when it wants particular content. It will insist on its technical specifications and open web protocols, but it will go beyond finding content on the open web. It will, in a sense, curate the content of its own indexes when that matches its strategic goals. Libraries tend toward cooperative arrangements with their data and resources, so this is a potential source of opportunity and a channel for libraries to expose their data.

However, the most effective way to gain visibility for any content is by following open web practices and making publicly available webpages that match Google's published best practices. The best information about how Google evaluates webpages is provided for traditional search results. Traditional search results are the list of websites and documents that appears in the middle of a traditional browser or mobile search application page. The space reserved for the traditional search results is one of three zones that make up the search results page on a full browser: sponsored links, traditional search results (central zone), and the Knowledge Card. Mobile results are different but share many of the same characteristics.

On a mobile device browser, the zones are different, but the principles for how content gets there are the same. There are also rules for the display of content in the other important parts of the Google search results page: the sponsored links and Knowledge Card. But first, it is important to understand how ranking of the traditional search results works.

The mechanics of managing results in the central zone begins with crawling and indexing and ends with page ranking. Crawling and indexing populate Google's indexes so it has words to search and links to display in results. PageRank determines how often a page has been linked to—this measure of a page's popularity is a measure of its usefulness. If lots of other webpages link to this page, it must be considered useful—perhaps even authoritative. Once Google has content in its indexes, it can compare searches to those indexes and determine what to rank in the results. To do this, the search engine compares the search to the indexes and asks somewhere around 200 technical "questions" of the page content in the indexes—these are interrogations of the indexes to determine which pages have the best results and how they should be ranked. The rules are constantly changing, and the full details of the rules are Google's most important trade secret, but Google tells all website designers that at least eight of the questions are central to the process and that they should take care to observe best practices in relation to these questions:

1. Is the page blocked? Has the webmaster put a block on the page so it can be accessed only through a browser directly and not by a crawler?
2. Does the page include videos and pictures? Multimedia content is considered good—this means it is a page that people are likely to want to stay on.
3. What is the word frequency on the page relative to the search? For example: How many times do the words *Noah's Ark* appear on the pages in the index? High word frequency on the page is a hint at relevance to the search.
4. Are the words in the search in the title section of the page? This is a technical detail of HTML writing—the TITLE section declares what a page is about. That's another hint at relevance.
5. Are the search words in the URL of pages? This is another hint. If the keywords *Noah's Ark* are right there in the URL, perhaps this entire website is about Noah's Ark; that's a hint at relevance to the search.
6. What about adjacency? Do repeats of the keywords appear close to each other? Another hint.
7. Are there synonyms for the words on the page? This hints at a deeper understanding of the topic and overall content quality.
8. What is the overall quality of the website? Here Google uses the term *spammy*. It wants to link to sites that offer what the user is looking for—whether that is buying or learning, it doesn't want to link to lists of other pages with no added value. In this area, Google might also look at how frequently a page is updated. But Google's staff warn content producers not to focus too much on this. Focus on overall quality of the content, and you will attract links and therefore value. However, a page that is updated infrequently—let's assume months or years between updates—will appear to be stale and of lower quality.[1]

Google uses the answers to these questions, and many more, to determine how useful a page is relative to the search. If the answer to many of the questions is yes, then the yes pages must be relevant. The strength of the yes (how closely the pages and the keywords match through the filter of the questions) is a measure of their relevance. But Google also uses its innovation in search: the PageRank. PageRank measures how many other sites are already linking to a page: that's the final hint at relevance and usefulness. Google's founders introduced this concept as a distinguishing feature of its search product. They developed the algorithm to create a measure of value of a page through the proxy of how often the page is linked to. In other words, how popular a page is—how many times other sites refer to it—is a measure of its usefulness. This is a key element of ranking and also has relevance in the second key element of the geography of the results page: the Knowledge Card.

Google's Knowledge Card

Because advertising is Google's chief source of revenue—users of the site don't pay money to use it, they pay with their time and exposure to advertisements placed on the results page—the company has turned to providing more and more content directly on the results pages. It is not providing just links to pages that might answer a question like, "What time does the Cincinnati Bengals game start today?" or "Where is the new Star Wars movie playing?" or "Star Wars show times?" It is providing the answers to those questions directly on the search results page. For some answers, it isn't necessary to click to the page or document—the answer is given directly in the Knowledge Card next to the traditional results. The usefulness of this Knowledge Card and its apparent durability (it has been on Google results pages for several years as of early 2016) indicates that it is worth understanding how to get content into this zone.

The Knowledge Card, sometimes called the Info-Card or Answer Panel, is the second of three important areas on the Google results pages. Of course,

there are rules for how content is selected for display there as well.

There is some debate about how website managers can influence the visibility of their resources within the Knowledge Card. Richard Wallis, Semantic Web expert, describes it this way: "To get your content into the answer panel, recognizable semantic properties will prove more fruitful and effective than simple words."[2] There is a lot in that statement, and it is useful to understand more about linked data and the Semantic Web before the full value of the statement is revealed.

First a review: Wallis is saying that following the rules of the Semantic Web improves your chances of getting your content into the Answer Panel. When reviewing the rules for relevance in the central search results zone, it was clear that page rank and then words—their placement, their markup, their frequency, and so on—were key to relevance and utility. In the case of the dynamically generated Answer Panel, a different set of rules is more important. The general guidelines for following the rules of the Semantic Web are

1. quality and breadth of the internal graph
2. quality of connections to the global graph
3. recognizable markup[3]

Quality of the internal graph relates to a Semantic Web principle that states that content should be described in terms of linked references to the things you are describing. That means any reference to a person, a place, an object, or an event should include a universal reference to that thing. This includes reference to the holding organization. This is typically a Uniform Resource Identifier (URI). As with traditional library authority control, a URI provides an unambiguous and repeatable way to represent something. A URI provides both the "address" of a data item and a description of the thing it identifies. And as it is used repeatedly, systems can develop trust that the URI is reliable—it points to a site with authority and trust in describing something. If the search engine can find the same identifier on multiple pages, then it can more efficiently determine that the page is about the same thing; this is useful when matching searches to pages; an unambiguous identifier is always better than trying to match text—it produces a more confident match. So quality of the internal graph is measured in how frequently the things on a webpage are described using links to authoritative sources instead of just text—even if that text is consistent. Using links provides a better score in conforming to the rules of the Semantic Web.

Quality of connections to the global graph is enhanced by the use of global identifiers for things. An organization or even group of organizations can invent their own identifiers for things, but using existing identifiers that are already used on the web (the global web) is the approach that the search engines reward. This is a familiar concept for librarians who have created a number of widely recognized schemes for consistent description: the Dewey Decimal System for a single term describing what a published thing is about, and the many national name authority files: the Library of Congress Name Authority File in the United States, the Integrated Authority File (GND) among the German-speaking countries, and the various name authority files from the French National Library (BnF) are all efforts by local communities to describe things in a consistent way.[4] The value of consistency was always a reduction in cost in cataloging and some benefit to the reader in consistency in indexing. Somebody has already done the hard work of determining how to spell an author's name, for example, or the town he was born in, or the degree she earned at a particular university. The benefit of consistency in indexing is manifest when the user has a better chance of finding something if there are cross-references to various forms of an author's name. Furthermore, the display of results is cleaner when the persons contributing to the work are recorded consistently.

The same principle applies on the Semantic Web, but the specific incentive is to use globally recognized identifiers when they exist. And since Semantic Web description is meant for machines and not humans, it is common to use multiple identifiers for a thing. In fact, multiple identifiers can be an advantage. As with synonyms in traditional relevance ranking, it shows that a website has a deep understanding of a thing. So collecting and using multiple identifiers for a thing strengthens the breadth and quality of the internal graph and indeed the global graph when global identifiers are used. The emphasis on connections to the global graph implies that there is value in multiple sites referring to things in the same way. Semantic Web experts would say this strengthens the nodes on the graph, but the plain language way to say it is to compare it to a chorus: the more voices singing the same words in the same key and at the same volume, the stronger the impact on the audience.

Wallis's third element is recognizable markup. *Markup* in Semantic Web jargon refers to adherence to the recommended vocabulary schemes to draw Semantic Web concepts into all websites. The global search engines Google, Yahoo, Yandex, and Bing agree that consistency in markup and Semantic Web principles make their crawling and indexing work easier. They are fierce competitors, but they have found a common interest in how content should be presented. They all recommend adherence to markup specified on their website schema.org. It encourages isolating the persons, organizations, objects, places, events, and other things being described and, where possible,

identifying them with global identifiers. Librarians have been very active in influencing schema.org, so the vocabulary better represents bibliographic items and the libraries that hold or offer them. By encouraging use of a de facto common vocabulary across well over 10 million sites, schema.org has introduced a broad consistency to the Semantic Web that has previously been lacking.

The dramatic increase in use of schema.org on the web hints that website developers believe it will help with indexing. It is also important to know that the Google Knowledge Card appears to draw from a set of reliable and durable sources that influence Google's own internal knowledge graph. There are many sources for Google Knowledge Card data, but DBpedia is frequently mentioned in this context. DBpedia derives its data in part from *Wikipedia*, and the direct management of the quality of that data is important for success in appearing in the Knowledge Card. Kenning Arlitsch, the Dean of Libraries at Montana State University, who has experimented deeply with managing the visibility of the library and its specialized collections, explains that DBpedia "tends to be the primary source from where Google gets is information for the Knowledge Card." Arlitsch says bluntly: "If you don't have an article in Wikipedia to draw into DBpedia, then you don't exist to Google."[5] The knowledge graph that libraries can influence directly is therefore an important part of the Semantic Web infrastructure and can't be ignored in the question of library visibility on the web.

Google's AdWords

The third zone on Google results pages is the sponsored links or AdWords. In this zone it is the business relationship with Google that determines placement. To explain how these results are displayed, Google says in plain language: "Google may be compensated by some of these providers."[6] Presumably Google's business development teams negotiate contracts with these providers and use data they provide to display results matching searches. It is reasonable to assume that all of the rules for the traditional results and the Answer Panel or Knowledge Card zones are used, but the additional factor of payment for placement is the final element that determines what is displayed in the sponsored links zone.

Libraries have an opportunity in that the rules of the web are well understood and there is an art and science around optimizing content for search engine uptake that is now more than a decade old. The challenge for libraries will be to apply those rules and change the many decades of practice in catalog and data management. With this understanding of how things are exposed on the web above, it is useful to review what users want from libraries.

Notes

1. Google, "How Search Works: Crawling & Indexing," accessed February 11, 2016, https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html.
2. Richard Wallis (Independent Structured Web Data Consultant), interviewed by Ted Fons by Skype, 23 October, 2015.
3. Ibid.
4. Ioannis Papadakis, Konstantinos Kyprianos, and Michalis Stefanidakis, "Linked Data URIs and Libraries: The Story So Far," *D-Lib Magazine* 21, no. 5/6 (May/June 2015), http://dx.doi.org/10.1045/may2015-papadakis.
5. Kenning Arlitsch and Patrick O'Brien, "Establishing Semantic Identity for Accurate Representation on the Web" (presentation, Coalition for Networked Information Fall 2014 Membership Meeting, Washington, DC, December 8–9, 2014).
6. Danny Sullivan, "Once Deemed Evil, Google Now Embraces 'Paid Inclusion,'" *Marketing Land*, May 30, 2012, http://marketingland.com/once-deemed-evil-google-now-embraces-paid-inclusion-13138.