White Paper: **A Brief History of Search**

Greg Kidd
Founder and CEO
3taps

April 11, 2010

This white paper discusses certain tools and systems people use to find information, and how these tools and systems have changed over time. In tracing the evolution of search, I hope to provide perspective on "diachronic search," the time-ordered search recently popularized through Twitter.

I divided this paper into nine sections: in the first, I discuss search techniques developed for books; next, I describe initial efforts at computational search, prior to development of the Internet; the third and forth sections provide an overview of the birth of the Internet and the World Wide Web, and how each was, to some extent, driven by a desire for better search capability; in the fifth section, I describe the development of authoritative search by Larry Page, and the subsequent success of Google; sections 6 and 7 provide an account of the birth of Web 2.0 and Twitter; in section 8 I discuss diachronic search and how it relates to Google's authoritative search model; finally, section 9 considers the opportunities in and obstacles to the development of Web 3.0, which has been described as the "semantic web."

1. Book Search

Prior to the invention of writing, the task of information retrieval was left to the ability of the human mind to remember facts and fiction as organized in each individual's memory. The tradition of oral history relied upon elders imparting the stories of their forefathers to the younger generations. When the first use of mnemonic symbols evolved from ideograms to alphabetic writing, a variety of storage mediums ranging from clay tablets to parchment were utilized to store subject material.

Proliferation of books, along with attention to their cataloging and conservation, developed during the Hellenistic period. The creation of libraries, and in particular the Egyptian-based Library of Alexandria founded at the beginning of the third century BC, and continuing into the 1st Century CE represented the pinnacle of the breadth and depth of human knowledge. The "corpus" of material there exceeded 500,000 scrolls and books, often meticulously hand copied from original work sourced from around the eastern and western world. For half a millennium after politics led to the destruction of this great library, copying and conserving books plodded along – hampered by the high cost and time necessary to make or duplicate a work.

Johannes Gutenberg's creation of the printing press changed that by enormously lowering the cost for creating books. Subsequent proliferation of books on all subjects during the Renaissance led to an ever greater need to organize the burgeoning mass of subject material. The invention in 1876 of the Dewey Decimal System resulted in a 10-class, 100-division, 1000-section hierarchical system for organizing books. Works in a library could be stored and searched for on shelves – a methodology still in use today in over 200,000 libraries in at least 135 countries. While the Dewey Decimal System has competition (aka the Library of Congress classification system), it remains an example of the state of the art in search for the realm of information stored in books.

2. Computational Search

In the realm of numbers and computational logic, the world made modest progress from the abacus in 2400 BC to clocks in medieval times and the beginnings of mechanical computational devices in the sixteenth and seventeenth century. In 1801, a loom for the creation of woven patterns used the first system of punched cards to achieve "programmability" and storage of data. In the late 1880s an American invented standardized punch cards to record tabulated data on individuals for the 1890 U.S. census. The use of mechanical counters for the cards allowed the census data to be tabulated months ahead of schedule and under budget – and served as the foundation for the company that became the core of IBM.

Breakthrough contributions in logic came in the first half of the 20th century with Alan Turing's work on the role of algorithms, and with John von Newman's work on electronically stored program architecture. The 1943 Colossus computer was purpose built to break the German's Enigma code, and the 1946 ENIAC was the first general purpose machine capable of being reprogrammed to solve a full range of computing problems. While ENIAC was designed to calculate artillery firing tables for the Army, its first use was in calculations for the hydrogen bomb. When computer hardware moved from reliance on analog vacuum tubes to transistor based machines, huge improvements in speed and reductions in cost to build and operate were made possible and continue along a path toward exponential improvement today. In parallel to these quantum leaps in computing power, have been similar gains in efficiency in data storage.

Pioneering groundwork in the search realm began in the 1970s at IBM with the move from using linked lists of free form records to using "normalized" tables of fixed length records that could be more efficiently organized and searched. A structured query language or SQL refers to method for efficiently inserting, updating, and *searching* large sets of data. Commercial application of these database concepts included LexisNexis for structured search of data in newspapers, magazines, and legal documents. These services required a subscription to a set of servers and terabytes of storage owned and hosted by a single company that meticulously collects and enters all the data in question from over 20,000 discrete sources.

3. The Internet

In contrast to a centralized assemblage of searchable data, the origins of distributed computing and search began with a research project to connect three disparate Defense Department computers at Cheyenne Mountain, the Pentagon, and SAC in 1962. A need for "inter-networking" was defined, at the core of which was the need to connect separate physical networks into one logical network. To make such a network survivable, the idea of dividing information transmitted into message-blocks of a standard/arbitrary size that could be routed across multiple routes (to avoid a single point of failure) was established and utilized in the first implementation of the ARPANET between UCLA and the Stanford Research Institute in 1969. Initially a number of different networks sprung up and the challenge became solving how to connect the different networks via an overall "internetwork protocol." The resulting protocol became known as TCP/IP and was largely defined by 1978 and permanently activated in 1983. By the late 1980s, the TCP/IP standard became global and supported a number of overlay technologies such as email systems, usenet bulletin boards, and file transfer protocols. The first commercial dialup ISP set up shop in 1989.

The Internet Engineering Task Force (IETF) represented government funded researchers and eventually non-government vendors in quarterly meetings concerning standards. Tim Berners-Lee, while working in 1989 as a researcher at the particle physics laboratory CERN, wrote a paper proposing a network based implementation of hypertext. Released as a concept for public use, he announced in a use group his aims for a "WorldWideWeb" project to "allow links to be made to any information anywhere." In 1994, Berners-Lee founded the World Wide Web Consortium (WWW) at MIT with support from the Defense Advanced Research Projects Agency (DARPA) and the European Commission. Berners-Lee made the Web available free, with no patent and no royalties due.

4. The World Wide Web

The standards in Berners-Lee's proposal led to the implementation of generalized graphical browsers by students at Helsinki University (Erwise), Berkeley (Viola) and the University of Illinois (Mosaic). The later was commercialized as Netscape under the leadership of Marc Andresen. The browsers could depict pages at a burgeoning number of addresses hosted as "websites." Search engines and web directories sprang up to keep track of pages so that people could find things. Initially, only website titles were searched and cataloged – much like the logic of the Dewey Decimal system for organizing books. But computational models Lycos and Alta Vista mechanically "crawled" the web pages of sites to determine what words were included and how those sites should be cataloged. In parallel, human judgment based taxonomies of web content also sprung up around 1994. Originally titled after Stanford student founders as "Jerry and David's Guide to the World Wide Web" the renamed Yahoo (for "Yet Another Hierarchical Officious Oracle") became a benchmark for organizing and highlighting web content. Each of the afore-mentioned sites morphed from their roots in web crawling and directory building to becoming web portals. As the number of websites proliferated and usage grew, some site owners began gaming the search engines' methodologies in order to draw visitors. As a result, the issue of assessing the relevancy of search results became more pressing.

5. Authoritative Search

In 1998, another Stanford graduate student name Larry Page published a patent on a "Method for Node Ranking in a Linked Database" (US 6,285,999 filed Jan 9, 1998 issued September 4, 2001) that established the principles for establishing the *relevance* and *authority* of web pages as a basis for web search. The simple concept is based on a democratic voting algorithm that tabulates PageRank based on how many incoming links a web page has. Each link is counted as a ballot among all the other pages on the World Wide Web about how important a page is. PageRank is determined recursively – meaning that a ballot from a page that itself has many incoming links is more heavily weighted than a page with fewer links. The number of ballot weighting is logarithmic scaled in a manner similar to the Richter Scale for earthquakes. Thus a page with ten times as many links as another may only be weighted twice as much due to the logarithmic scaling effect. Larry had the patent assigned to Stanford and then co-founded Google, and exclusively licensed the patent. The new basis for authority-based search of web pages outflanked the models used by Lycos, Alta Vista and Yahoo – eventually knocking those players out of the core search business.

Google's growth and success occurred largely after the Web 1.0 dot com boom-bust. Initially, it was not clear what the business proposition was for pure search of all the Web and portal sites that were each tripping over themselves to garner traffic – with or without a business model for monetizing that traffic. Google's PageRank algorithm clearly garnered usage as a superior tool for determining which search results should be placed at the top of a results listing – but it was the implementation of the paid-for ability to push sponsored links to the top and side of the list (AdWords) that mated science with monetization. By allowing firms seeking click-through traffic to their site to bid for inclusion and position in search results, Google laid the foundation for a revolution in how advertising and marketing campaigns in the new media could and should work. In contrast to the scattershot approach of banner advertising transferred from old media (newspapers, TV, radio) to the Web 1.0, Google offered a targeted "pay only for results" approach that was relatively non-obtrusive for its users and performance efficient for its advertisers. With AdWords, Google proved that search was not just a nerd field for computational linguists, but also a lucrative – perhaps the most lucrative – business model in the world. Google's growth and market capitalization soon reflected its stature as the most successful internet company of all time – all built on a superior understanding and execution of "authoritative search."

Google's success was built in the wake of dot com disasters with no business model (Webvan, Pets.com), challenged business models (Yahoo, AOL), and those with proven transformative business models (Amazon and Ebay and Craigslist) – all of which have roots in the Web 1.0 era. The more successful of these models all tend to have a large amount of user created content – basically postings of goods and services for sale (either directly in the case of eBay and Craigslist, or indirectly via publishers in the case of Amazon). User content is cheap to create relative to proprietary content (as AOL and other proprietary portals have found to their dismay). User ratings of buyer/seller experiences also represent another form of determining relevance and authority in the realm of the exchange space – something that Google's page rank system doesn't account for.

While Google search excels at determining that eBay may be an authoritative place to go for auctions of goods, it doesn't tell you which post is particularly relevant versus another one for a similar search term. In fact, Google's algorithm for counting ballots will tend to the find the oldest and most stale posting for a given item on a given site – exactly the opposite outcome to the principle of "best execution" associated with efficient and effective market places. Nor have search robots been particularly effective at ferreting out best execution of prices, supply, and demand across diverse data structures trapped in various site silos across the Web.

6. The Web 2.0

The Web 2.0, which first gained traction as a concept in 2004 when O'Reilly Media hosted a conference on the topic, presents a set of unique challenges and opportunities in the web search space. Whereas the Web 1.0 treated users as passive retrievers of information presented to them, Web 2.0 anticipates a world in which users create and change the Web with their own content via hosted services, social networks, wikis, blogs, and video sharing sites. While the Web 1.0 had its share of user contributed content via classified sites, job boards, and personals, the Web 2.0 entailed much more dynamic and unstructured content as Facebook and Twitter have popularized.

When Yahoo acted in a Web 1.0 mindset and bought the personal website company GeoCities for $3.6B, it was the third most visited site on the Web (behind AOL and Yahoo). When Yahoo changed the terms of use (TOU) and claimed ownership of all posted content from its users, it started a maelstrom of protest from its members that the site never recovered from – resulting in total destruction of the user base within a decade.

Conversely, Web 2.0 companies have embraced openness – both for its users, and to developers through liberal use of application programming interfaces (APIs) and inbound and outbound syndication of data via RSS, RDF, Atom, and XML standards. Popular users in the new Web 2.0 companies develop their own relevance and authority amongst friends, peers, and any "follower" in the greater public. For the first time since the emergence and dominance of Google Page Ranked based search, an alternate method for determining relevance on the Web – i.e. a referral from a trusted source on Facebook or Twitter, is beginning to garner as much traction for sourcing click-throughs to destinations on the Web. The disruptive effect of social media on old media is already a well discussed topic on the Web – but what has been less well documented is the disruption these networks are causing to the formerly nascent but lucrative established world of authoritative search.

7. Twitter

While Facebook has garnered the most Web 2.0 media attention for its large user base and trumping of the less agile MySpace, the real disruptor in the web search space is Twitter. Originally dismissed by Google CEO's Eric Schmidt as a "poor man's email" at a Morgan Stanley technology conference in 2009, Google subsequently agreed to pay for access to Twitter's firehose of status updates (tweets) for integration into its own search engine results in 2010. Google was forced to face a choice in a competitive market where Microsoft also acquired, at the same time and on similar terms, the same stream of *postings* from the poor man's email system. Poor or not, the stream of postings provided by Twitter represented the closest thing to the "pulse of the planet" on breaking news. While the sources from Twitter may or may not be individually authoritative, the collective wisdom of status updates across a spectrum of users at the moment of a new event (such as an earthquake or revolution) trumped any other source for immediacy and discernable trust and authority. Coupled with the inclusion of updates from noted users with huge follower bases (such as Oprah, Lance Armstrong, and in the run up to the election, Obama), the absence of such updates would represent a huge hole in any search engine purporting to be authoritative.

Twitter postings are unlikely to be authoritative relative to any static knowledge already in a book or on the Web – i.e. information that has already been collected, tabulated, analyzed, and repackaged into a blog or static web page. For information where state or status is already known, Twitter postings have no comparative advantage as far as the creation of information is concerned. The comparative advantage occurs in two realms:

i. the announcement of *change* in state or status;

ii. the authoritative conveyance by a party posting to a party following of a link to the full/best information on the change in state or status.

If one thinks of books as eternal markers of the state of the world (locked in print at the time of publishing), then web pages are the eternally updatable but still (at any point in time) static markers of the state of the world. What that leaves for Twitter as a status update system is a role as the *derivative* of state and status – i.e. a marker for the change in state. The body of all Tweets then is a corpus of markers of change in state – sort of what the ticker tape of changes in stock prices and the number of shares bought and sold is to a stock market exchange. The knowledge Google holds regarding page rank is indeed valuable, but it is valuable in a historical way similar to what Value Line knows and publishes about the historical prices and volumes of trades for any stock that it covers. But what a Bloomberg terminal provides to a trader is real time information on changes in prices and volumes *at the market*. And while there is value to knowing the historical status of price and volume information relating to an exchange, best execution and decisioning is dependent on a current understating of the level, direction, and pace of real time market information. Or to put it another way, when it comes to best execution and decision making, knowing the actual market (regardless of whether the market is right or wrong on Monday morning quarterback basis) is the best information one could ask for. Operating without knowing what's happening on the margin – i.e. what the change in state happens to be – puts a trader at a fundamental disadvantage to peers operating with the best information available.

Google search without the integration of Twitter feeds is fundamentally vulnerable to being trumped by Twitter search for breaking news. The conundrum of Google being forced to pay for access to this data stream is testimony not to any flaw in the PageRank methodology, but simply an acknowledgement that the Web 2.0 reality has created a parallel universe for determining relevance and authority for breaking (rather than established) knowledge. Further analysis of the situation reveals that Twitter trumps Google in three ways that are disruptive to the prior order of determining search relevance on the Internet:

i) Most of the Web is about "state" – which is fixed at any point in time (though less so than a library of books that are truly fixed in print). Static knowledge is valuable, but in a world of rapid change, *the most relevant and actionable information for some types of decisions and actions relates to "change in state."* For all its power, Google only catalogs the Web at any point of time, with no method for comparing, ranking the importance, and portraying the relevance of change in state;

ii) The TCP/IP and WWW standards that have allowed the Web and Google's cataloging of the same to achieve a vast scope of coverage is actually trumped by an even more ubiquitous and far reaching standard supporting SMS and mobile telephones worldwide. *Because Twitter is built to this lower (and arguably more crude) common denominator standard, it is supported more ubiquitously as a platform for syndicating (in- and out-bound) data – and in particular brief bursts of status update data*. While the format is limited in the type of media supported and length of info allowed, the ability to embed links enables the richness of the Web to still be tapped by

Twitter while still maintaining a scope of user access that trumps what Google can directly evaluate with its PageRank search methodology;

iii) Twitter has been (and still is in some quarters) mistakenly classified as a social network (and an inferior one at that in comparison to Facebook). Particular criticism lodged at Twitter relative to other Web 2.0 offerings is that i) the number or active members sending out tweets is low; ii) there is an imbalance of followers to members in that a few members have many followers and in turn follow few if any people; iii) the Twitter website is not very sticky and many users don't even bother to come to the site. All of the observations are true, and they all miss the point that *Twitter is only an incidental social network, but that it is first and foremost an alternative communication channel to the internet itself.* What Twitter is, is the most efficient means for collection and syndication of posts on changes in state, with a self-regulating means for eliminating the vagaries of spam by limiting transmission only to parties that actually opt in to follow posting updates – and recognizing and respecting the means under which they agree to receive that information. The foundation for this self regulating system is the ability to follow (listen to) any source of information and, at least as important, the *ability to stop following when a source no longer is relevant or authoritative enough to be worth listening to*. Viewed in the context of a leveraged communication network, the very attributes of asymmetry of usage between peers indicates the strength of the model as an exponential, yet still opt in, system for getting the word out – but only to those who freely decide they want to receive it. Such a balance between the sender and receiver of status update information is a baked in assurance that conveyed information will be both relevant and authoritative (from the point of view of the recipient themselves).

Just as Google's radical innovation to web search was to introduce a ballot system for page ranking, Twitter has contributed its own ballot system – but for postings. Whereas Google counts links and links alone in determining the order of its search results, Twitter is always keeping track of followers (aka the social graph) and dutifully posting any and all posts for someone or something followed based on reverse chronological order. A simple matching of a search request to a user name or a searched for word results in a reverse chronological real time search result of the most recent posts and new posts as they come in. Whereas Google focuses on periodic updated counts of links, Twitter endeavors to keep its social graph updated in an even more real time mode while churning out a firehose (and various filtered garden hoses) of its “diachronic corpus.”

8. Diachronic Search

Diachronic corpus means the change in the use of words in a body of work – in this case the corpus is all status updates running through the Twitter utility. Diachronic search – implies that more recent information, all other things being equal, is more valuable that the information that came before it. While any particular new post may be untrusted, a compensating ballot value can be determined by looking at who posted it and the trust placed in that source. If that source is not directly known and trusted by the viewer, one can conceivably look at the trust placed in the followers of that party – and so on in an infinitely recursive “degrees of separation” analysis of the social graph of who is following who.

Mathematically, the calculation of “inbound links” conceptualized in Larry Page’s 1998 patent application can be refactored and applied to follower links in social networks to determine a “PostingRank” of how much authority or trust can be placed on a particular posting flowing from Twitter. Furthermore, because the links are related to people rather than pages, the application of a PostingRank concept can be applied not just within Twitter and its corpus of postings, but to any activity on the Web or in the non-virtual world conducted between parties that have a reputation established via a follower network in Twitter. For example, the same trust calculation could be used to calculate a trust score for the confidence that should be placed in an exchange transaction between two parties with established and validated Twitter reputational identities. Or when eBay like ratings of an exchange transaction between two parties is given (thumbs up or thumbs down), the weightings put on the references could weight (on an log exponential basis) those ratings so that they could be summed to represent the accumulated social capital of the parties involved. One could call this the 'Oprah effect' of recognizing the weight of her endorsement of Obama over Clinton in the 2008 presidential campaign. Oprah currently has 3.3 million Twitter followers (while following 19 people herself – including the Shaq). The PostingRank is simply a mathematical way of dynamically reflecting how to weight search results that reflect the power of postings from someone who is so listened to – both on and off the Web.

Not surprisingly, Twitter has asked itself how it can make money on its search function without alienating users. Taking a page from Google, Twitter has announced that it will also accept being paid to have certain posts appear as sponsored posts when certain search terms are entered and an advertiser has paid to "post on demand" when those terms are entered by a user. Both Google and Twitter have eschewed the use of banner advertising and argue that their sponsored links and pay-for-click-through models don't really amount to anything more that a smart pay-for-performance integrated marketing platform for organizations seeking targeted information dissemination. Increasingly, they have good company in other vertical segments. Whereas Craigslist still bases its revenue model on a Web 1.0 construct of charging people to post, a content aggregator like Indeed, working in a vertical segment for job postings, believes that as much or more money can be made by playing the sponsored link and click through game. While its true that Craigslist may be profitable today, and clearly acted as a disruptor of prior business models in the print classified space, the Indeeds of the world may be delivering a stronger and more defensible value proposition going forward based on applying a more sophisticated search algorithm to determine just which job postings should be portrayed at the top of the page for a particular user search.

The differences between Google's authoritative search and Twitter's diachronic search do not imply that the two firms need be at war over the relevance of their search results. As recent business arrangements suggest, there's a divide up the search space work attitude that suggests a "horses for courses" approach to using the best method for a particular situation. Twitter doesn't have the interest or resources to do what Google does in the dominant search space, and Google is willing to cede (for now) Twitter's dominance in the status update space and integrate those results (with their own special interface on the results page) when it feels a topic searched for is particular ripe for presenting real time search updates. And Google is by no means dependent on Twitter as a sole source for real time updates – a fact that is reflected by their taking a feed from many other prominent non-Twitter sources where ever it can. It's just that Twitter has grown so fast and so publicly in the breaking news update space that it has become a de facto standard (whereas Facebook still dominates in the private space of updates for friends).

9. The Semantic Web

The Web 3.0 has been called the "semantic web" and is expected to constitute a system that better "understands" what people want to request and do with all the content available on the internet. Tim Berners-Lee has again spoken with a voice about the importance of "exchange" – be it for data, information, and knowledge. Whereas much of the valuable information today on the Web is "in" various websites and can potentially be found via search to find those sites, and/or search within those sites – the implication here is that such access is not good enough. What the Web 3.0 anticipates is that a framework exists that allows data to be shared across applications, enterprises, and community boundaries. To achieve this end, data needs to be liberated from silos that hold it so that the repositories for such data don't act as walls to discovery or use of the same said data.

Searching is just the beginning hurdle to be overcome for getting access to postings that may lay deep within a site that collects and provides access to specialized sets of data. Users are likely to want to converse about such data and possibly conduct exchanges – possibly monetary exchanges. The ability to do so, if constrained by sites and silos that put the interests of control and ownership of the content ahead of the end users (in the way that Yahoo did with GeoCities) puts the potential of the semantic web at risk.

From a user and business point of view, the semantic web is all about removing restrictions on the free flow of information that undermines the efficiency and ubiquity of search. Two sets of challenges sit squarely in the way of the semantic web. The first is technical and pertains to getting all the data to be searched and acted upon into standardized location and category taxonomies that allow standardized searching and processing at the per post level across (rather than within) the various websites that data sit in. Just as the IETF had to develop standards that allowed computers and computer networks to be interoperable, so too are standards needed for data housed in various proprietary databases sitting on disparate servers around the world. End users would benefit if such data existed in what amounted to a single virtual pot of all data – regardless of whether it was originally sourced via Craigslist, eBay, Match.com, or Monster.com. Even with XML tagging standards to allow easy translation of some common fields for a particular domain of data in classifieds, there is still no guarantee that the same fields, whatever they are called, or how they are formatted, are going to be used in one website silo versus another. Common denominators must be found between data sources for search to be efficient and not burdened by too many false negatives and positives about what constitutes “like data” for a given search.

A good example of progress in the arena of “light” standardization of data tags necessary to support the semantic web is the adoption of hash tagging conventions within Twitter. Type the terms #job #engineering into Twitter search and you will be returned mostly postings with links to new job offerings in the engineering arena. If such data is further tagged with a recognized location indicator such as #SFO for San Francisco, the same search can become location aware. As yet there are no established top down standards for hash tags – thus the conventional uses for terms are being developed by the “mob of convention.” Such bottom up mob standard setting is not at all inconsistent with the spirit or the democratization of the Web 3.0 space.

A second obstacle to the uptake of the semantic Web 3.0 potential is outright objections by existing Web 1.0 and Web 2.0 organizations that will seek to preserve their status and market position by blocking the disruptions that the new order implies. While some websites will see the inbound and outbound syndication of their data as an opportunity for new growth of sources of data and ways of usage, others will see a direct threat to their model of shepherding users to their sites, banner ads, and user membership based models of access. The thought that their postings, even if listed already on the public internet rather than sitting behind a firewall in a private exchange, might become viewable and actionable outside the confines of their administered website, is threatening. The main method of expressing this fear and assertively attempting to protect the status quo is an aggressive terms of use that penalizes usages of their data that might enable the semantic web. Typical defensive TOU's attempt to make use of three legal defenses to ward off challenges to Web 1.0 and even some Web 2.0 ways of thinking:

i. Claim copyright or a license to control the content in a copyright like manner over content created by users but hosted in a particular silo based website. Sue 3rd parties that reuse that content verbatim and possibly sue or threaten to sue 3rd parties that even have just an analysis or summary of the data in question;

ii. Claim that access to the data, even if not copyrighted or protected as proprietary in some manner, is achieved via some crawling or scraping method that represents either an undue resource burden on the site, or an actual computer trespass. Sue 3rd parties that access the data via these means as if they are performing authorized access and therefore a potential criminal trespass;

iii. Claim a competitive right to exclude parties deemed to be in a competitive position to data and means that any other non-competitor arguably has unfettered rights to. This amounts to a "separate and unequal" clause whereby particular 3rd party users are singled out on a knock list of denied access that the general populace of users would not or could not be excluded from. Sue 3rd parties that still attempt to access data that all other comers might have access to just on the grounds of being able to execute the prerogative of denial of access (i.e. "we reserve the right to refuse service to anyone).

Just the threat of a cease and desist order and the cost of a legal defense by a startup Web 3.0 company in a fight against a well established Web 1.0 or 2.0 company can be enough to smother semantic web innovation in the cradle.

Craigslist serves as the poster child for websites clinging to Web 1.0 strategies. It aggressively sues innovators in the semantic realm who seek to create mashups of postings that cut across geographies or include thumbnails for postings with pictures. At the same time, it has been slow to innovate, generally refusing to add improved search features to its own site, or by 'freeing' the data for search on the Web. The idea of dis-intermediating Craigslist data from the Craigslist site silo is anathema to Craigslist.

Here we see two distinct models at work: the 3.0-compatible model suggested by the cooperation of Twitter and Google, and the 1.0-based silo approach of Craigslist. Thus a search for the term "Porsche 912" will likely find few Craigslist postings in Google search results – even though Craigslist has the largest supply of local posting for Porsche 912s for sale in the country (more than even eBay). Going further, and adding the word Craigslist to the search as in "Porsche 912 Craigslist" will likely turn up some matches. But because of the authoritative algorithm for Google's PageRank methodology – the most "authoritative" match on a classified posting will likely be the oldest one that 3rd parties have established inbound links to. Unfortunately, those links indicate an ad that is likely stale – meaning that the item is already sold or the ad is expired. In short, in this case, the most authoritative page is the least rather than most relevant. Here is a situation where diachronic search performed in Twitter would be by far the better search mechanism.

Unfortunately, Twitter's diachronic search capability, which would work so well on "discovering" the newest posting of a Porsche 912 for sale, works only on the data that Twitter has access to see. That would be fine if new postings to Craigslist were recognized in Twitter as status updates. In fact all exchange transaction postings could be thought of in the exact same way – as potential status updates to Twitter (or whatever the recognized semantic equivalent is that acts as a repository for postings that relate to exchange activity). But Twitter is passive – it only indexes those updates it receives. And despite the fact that such syndication might bring more traffic to the posts that Craigslist hosts on behalf of its users, it is not apt to play nicely in the sandbox with anyone that conveys its content (or even a search summarization of its content) outside of its own in house silo search offering on its own website.

While Craigslist is an extreme case of a Web 1.0 company that goes out of its way to make it difficult for Web 3.0 companies to gain traction, it is indicative of one end of the spectrum of possible reactions to the next evolution of search. From the smallest to the largest source of exchange postings, the challenge of removing friction to enable data to be liberated is imperative. That friction may be purely technical and include the following hurdles:

- the mechanism to grab the data, either via a search API or RSS feed
- the mechanism to translate the data into a common location and category taxonomy
- the mechanism to normalize the data into a common set of fields despite originating from disparate sources with excess, missing or different sized fields
- the mechanism to index all of the content in the posting, including free form words and contextual tags that apply to particular category domains
- the mechanism to summarize the posting and create a header and tiny URL link back to the original hosted content
- the mechanism to post the data in real time, without latency, and with assurance that the entered data is appropriate and does not violate legal standards for content
- the mechanism to serve the data to 3rd party content users via a search API
- the mechanism to notify 3rd party content users via a firehose or garden hose of posts that match pre-defined filter criteria for syndication
- the mechanism to host conversations between seekers and providers that match over particular posts
- the mechanism to support commerce based transactions for postings that have monetary exchange value
- the mechanism to allow users to rate the quality of exchanges and by so develop reputational capital within the exchange community

- the mechanism to perform clearing and settlement between seekers and providers that need an objective 3rd party escrow for either payment and/or presentment of goods, services, and information

The challenge is that all of the hurdles exist for data to be syndicated from or to the smallest content providers and content users and the largest. Craigslist and Twitter represent the largest examples in the content provider and content user categories, but they are not conceptually or mechanically much more complicated than smaller brethren. That's good for the semantic web, in that if proof of concept can be shown between the largest content providers and users, then the same can be true for any smaller parties. The challenge is how to efficiently scale that roll-out, so that the data on the Web starts out in an easy to syndicate format, rather than achieves that status through "design after the fact" of silo origination. That's the challenge of the semantic web, and it's a challenge and opportunity that 3taps is dedicating itself to solving.

The W3C site answers its own question in the FAQ section of its website. The rhetorical question asked is "What is the killer application for the semantic web?" The answer given is:

'the integration of currently unbound and independent "silos" of data in a coherent application is currently a good candidate'

We agree.