# Local Search: The Internet *Is* the Yellow Pages

*The proposed Internet-Derived Yellow Pages aggregate, annotate, and certify Web content for use in geographically oriented searching. The IDYP provides a framework for combining Internet-derived content with the trust and fairness that characterize the printed Yellow Pages, still the predominant source of consumer-oriented local information.*

*Marty Himmelstein*
Long Hill Consulting, LLC

Every day, millions of people use their local newspapers, classified ad circulars, Yellow Pages directories, regional magazines, and the Internet to find information pertaining to the activities of daily life: nearby places, local merchants and services, items for sale, and happenings about town.

The Internet is not meeting its potential for delivering this type of geographically oriented information. Sometimes the information that people seek is on the Internet, but the tools for locating it are inadequate. In other cases, our industry has not developed the counterparts to replace traditional delivery methods such as the printed Yellow Pages.

The trends that point to the rapid growth of geographically oriented search, known as *local search*, are unmistakable. The most important predictor of the intensity of an individual's Internet usage is the availability of a broadband connection. As of early 2004, 55 percent of all US adult Internet users had access to such a connection.[1] Further, the number of adult Americans who had broadband Internet connections at home increased 60 percent from the same time in 2003, to 24 percent.

Broadband access makes the Internet a pervasive, "always-on information appliance."[2] People with high-speed access do more things on the Internet, and they do them more frequently. The Internet has always been used to support local activities, ranging from Yellow Pages searches, mapping, and vacation planning to researching products prior to purchasing them in a nearby store. Ubiquitous broadband access will serve to increase users' expectations for better support for all types of location-based computing.

On the search side, a market study of 5,000 online shoppers conducted by TKG and Bizrate.com found that 25 percent of the responders' searches were for merchants "near my home or work."[3] A recent Forrester Group study found that 65 percent of online shoppers researched a product online before purchasing it offline.[4]

On the content side, at least 20 percent of Web pages include one or more easily recognizable and unambiguous geographic identifiers, such as a postal address. Many of these Web pages have locally relevant content; Web authors don't put addresses on pages haphazardly. This content is already on the Internet despite the lack of an overarching mechanism for accessing it.

On the revenue side, US businesses spend $22 billion annually on local advertising, $14 billion of which is for the Yellow Pages, but only a sliver is for the Internet. Greg Sterling, managing editor for the Kelsey Group, a research firm that provides Yellow Pages metrics, puts the upper limit of advertisers worldwide who purchase paid search slots on the Internet at 250,000, but few of the slots are for

Published by the IEEE Computer Society

local content.[5] Contrast this with the more than 12 million small and medium businesses (SMBs) in the US, and another 20 million or so in other developed countries. The predominant market for SMBs is local: 60 percent of businesses in the US report that 75 percent of their customers come from within a 50-mile radius.[5]

## LOCAL SEARCH TODAY AND TOMORROW

Local search today is discussed in the context of paid listings—the advertisements that appear near the algorithmically computed, or natural, results search engines return in response to user queries. However, paid listings and their variants are not the bedrock upon which local search will be built. To see why, we only need to examine the original and still predominant local search tool, the printed Yellow Pages.

The Yellow Pages have many shortcomings, but they also have two virtues that are indispensable for local search: They are both trustworthy and inclusive—they contain at least minimal information on all businesses.

Paid listings do not provide the infrastructure for replicating these core Yellow Pages virtues—in fact, the value of paid listings is that they are the opposite of fair. Rather, to reach the widest audience, paid listings require a stratum of YP-like data beneath them, and the richer that stratum is the better.

The challenge for the local search community, then, is to facilitate the creation of this stratum of data. It must create better ways to collect and disseminate geographically oriented information about the activities of daily life. To meet this challenge, local search must supplant both the printed Yellow Pages and the current generation of Internet Yellow Pages (IYP)—a transplanted direct-mail mailing list—as a means for gathering and presenting consumer-oriented business information.

In ways that are readily evident, the Internet can furnish richer content than the Yellow Pages, but it cannot yet duplicate its orderliness and fairness. And fairness is the crucible by which local search will be judged. If users don't trust local search, it won't matter how much better than the Yellow Pages it is. People won't depend on it.

People use the Yellow Pages occasionally, but they are involved in local activities continually. It is therefore natural for local search to reflect the range of activities in which people participate. For example, much of our local activity has a temporal component. The Internet has the potential to provide access to transient local information more efficiently than older distribution mediums. A definition of local search that encompasses its temporal, commercial, and noncommercial aspects is that "local search tells me what is located within 100 miles from here and what is happening within 100 miles from here."

Broadly speaking, there are two sources of local content on the Internet. *Offline-derived* local content originates from other, usually older, sources, but is distributed on the Internet. The IYP is the primary source of offline-derived local content on the Internet.

**The Internet-Derived Yellow Pages uses the Internet for both content distribution and content aggregation.**

*Internet-derived* local content is gathered directly from the Internet. While many searches return pages with local content, to date only a few systems have attempted to gather and present content that is specifically relevant for local search. Geosearch, a joint project between Vicinity and Northern Light, was the first large-scale effort to derive local content directly from the Internet.

Currently, content aggregators, such as the various city and vacation guides that abound on the Internet, provide some of the best local content. These aggregators have good information for popular categories, such as lodging and restaurants, and they rely on the IYP to fill gaps in their coverage. While they are good sources for some types of content, they do not provide a mechanism for replacing either the print or Internet Yellow Pages.

The Internet-Derived Yellow Pages provide a framework for local search that incorporates the trustworthiness associated with the Yellow Pages without jettisoning the potential for distributed, unencumbered content creation that is the Internet's inherent strength. The IDYP uses the Internet for both content distribution and content aggregation. Aggregation is a more significant challenge than distribution, but one that is not adequately addressed by the local search community.

The IDYP's ken is wider than commerce, but local search's first requirement is to be a better Yellow Pages.

## GEOSPATIAL PROXIMITY SEARCHING

All varieties of local search require the ability to find information associated with locations within a given distance of a specified search center, known as *geospatial proximity searching*. Preparing data sources for proximity searching requires several steps.

The first step is to locate text that the geoenabled application can map to a physical location. This step is easy for the IYP because it is a simple structured database with defined meanings for each field. For Internet-derived content, the problem is trickier because text with geographic significance can be anywhere on a page.

The second step is to transform a location's textual designation into physical coordinates on Earth's surface. As the "Detecting Geographic Content in Text Documents" sidebar describes, the topic of detecting geographic content within text documents has generated interest in both the commercial and research sectors.

**Geosearch transforms location information in text documents into a form that search engines can use for efficient proximity searching.**

In the developed world, a street or postal address is the most common way to refer to a location, particularly for local search. Geocoding applications attempt to resolve a group of tokens into a pair of geographic coordinates, usually expressed as latitude and longitude. Along with each pair of coordinates, a geocoder also returns a value that represents the quality of the returned geocode. The best geocodes are accurate to within a few meters; less specific coordinates usually refer to the centroid of a larger region. Geocoding databases are large and dynamic, like the street networks they represent.

Efficiently processing proximity queries against large data sets, such as a nationwide business directory of 14 million businesses, or the Internet, requires *spatial access methods*. The basic idea of SAMs is to map a two-dimensional (or $n$-dimensional) coordinate system—in this case latitude and longitude—onto a single-dimensional coordinate system. By doing so, a region on Earth's surface can be denoted with a single attribute, a *spatial key*, instead of the four attributes that are necessary to describe a bounding rectangle: the $x$ and $y$ coordinates of the rectangle's lower left and upper right corners.

Spatial keys are computed as the data set to be geoenabled is being built. In the case of a business directory, the spatial key for each business is stored as an additional field along with other fields for the business. If a search application is indexing unstructured text, it adds the spatial keys as additional terms to the index it builds for the page. Unstructured text, such as Web pages, could require several spatial keys because they may contain several addresses.

At search time, to determine the set of businesses or Web pages that satisfy a proximity query, the search application maps the user's search center and desired search radius to a set of spatial keys that cover the area to be searched. It then adds these spatial keys to the user's nongeographic query terms so that they can be compared to the precomputed spatial keys stored with the dataset being searched.

The search application refers to the user's nongeographic terms to determine which data items within the radius match the user's main search criteria. Proximity searching algorithms can order results by distance, so results closer to the search center are listed before those farther away. Ordering is routine for IYP applications, but can be problematic for Web pages because of the potentially large number of pages that may need to be sorted. An example of a paraphrased Geosearch query is: "Return Web pages that are about hot-air balloon rides and which contain postal addresses or telephone numbers within 100 miles of 10 Main Street, Poughkeepsie, NY."

## INTERNET-DERIVED LOCAL CONTENT

In 1998, a research group at Vicinity developed a prototype system to geoenable Web content. Vicinity modified the spatial access methods it developed for its IYP and business locator products to work as a software component in conjunction with search applications. In 1999, Vicinity teamed up with Northern Light to broaden its experiment to include the general Web corpus. Microsoft purchased Vicinity in December 2002.

Geosearch was publicly available from April 2000 until March 2002 from both Northern Light and Vicinity's MapBlast property. During this time, Geosearch processed about 300 million distinct Web pages.

The experience with Geosearch provided the basis for two observations:

- the Internet is already a rich source of local content, and
- local information on the Internet possesses certain characteristics that simplify the job of aggregating it.

The basic idea of Geosearch is that it transforms location information in text documents into a form that search engines can use for efficient proximity searching. Its first step is to scan documents to recognize text patterns that represent geocodable entities.

Geosearch avoids semantic text analysis, preferring to leave the determination of a document's subject matter to the information analysis algorithms of the search application with which it works. Geosearch relies on the fact that a significant portion of the content that is valuable for local search

## Detecting Geographic Content in Text Documents

The topic of detecting geographic content within text documents has generated interest in both the commercial and research sectors.

### Commercially available products

Google Local (local.google.com) scans Web pages for US and Canadian addresses and North American Numbering Plan telephone numbers. Whereas Geosearch used location purely as a filter, Google adds an extra step of trying to correlate the address information on Web pages with IYP data.

Most local search offerings combine IYP data with more in-depth content from vertical content aggregators, but so far, Google is the only search engine that uses the Geosearch approach of deriving local content directly from Web pages. One sure way to determine whether a search product obtains local content directly from the Internet is to do an idiosyncratic search for which there is unlikely to be any IYP data. For example, both Geosearch and Google Local return results for "worm composting in Thetford, VT"—others do not.

MetaCarta's Geographic Text Search (www.metacarta.com) is a commercially available product that uses a place-name directory in combination with context-based analysis to determine the presence of geographic content. It will, for example, assign a location to the phrase "three miles south of Kandahar." GTS is appropriate for corpora that might have geographic content but not the obvious markers of postal addresses or telephone numbers.

### Research

Content-based searching for location information requires identifying tokens that might have a geographic meaning. Systems that use place-name directories, called *gazetteers*, need to check the gazetteer for every token in a document. A token that is in the gazetteer must then be disambiguated to see if it really represents a location, and if so, which one. This process can be costly.

Systems based on standardized addresses typically look first for postal codes. Tokens that look like postal codes are rare, so few trigger additional work. Then, since the sequence of tokens in an address is rigidly constrained, it is not difficult to determine if a potential postal code is in fact part of an address. Efficiency might not be a concern for some document collections, but it is if the collection is the Internet.

Web-A-Where,[1] a gazetteer-based system that associates geography with Web content, uses several techniques to resolve ambiguities. Ambiguities are classified as geo/geo (Paris, France or Paris, Texas) or geo/non-geo (Berlin, Germany or Isaiah Berlin). The system also assigns a geographic focus to each page—a locality that a page is in some way "about."

Junyan Ding and coauthors[2] analyzed the geographic distribution of hyperlinks to a resource to determine its geographic scope. As expected, their analysis showed that *The New York Times* has a nationwide geographic scope. However, so does the *San Jose Mercury News* because readers across the country follow this California newspaper's technology reports. These authors also estimated a resource's geographic scope by using a gazetteer to examine its content.

In contrast to Ding and coauthors, Geosearch and Google Local rely on a user-centric approach to determine geographic scope because they allow users to specify the search radius of a query.

Kevin McCurley[3] discussed using addresses, postal codes, and telephone numbers to discover geographic context. Remco Bouckaert[4] demonstrated the potential of using the low-level structure of proximate tokens, such as postal addresses, to perform information extraction tasks.

### Organizing Web content for local search

With the exception of the work by Dan Bricklin,[5] relatively little has been written about organizing existing Web content for local search. Bricklin proposed the small and medium business metadata (SMBmeta) initiative as a way for enterprises to present essential information about themselves consistently on the Web. The idea is to create an XML file at the top level of a domain that contains basic information about the enterprise. Since SMBmeta files have a consistent location, name, and structure across Web sites, search applications can easily find and interpret the files.

In a perfect virtual world—a Web presence for all businesses, the willing participation of search engines to promulgate the use of metadata standards, and no spam—the original SMBmeta initiative would offer a simple way to disseminate information about local businesses.

In lieu of this, Bricklin proposed the SMBmeta ecosystem, which sketches some control mechanisms that are similar to IDYP's trusted authorities. Upon request, a registry returns a list of the domains it knows about that have SMBmeta data. A proxy maintains the equivalent of the smbmeta.xml file for domains that do not have their own files. An affirmation authority performs the policing functions.

Meeting the IDYP goal of creating an Internet version of the printed Yellow Pages requires leveraging the political and organizational infrastructure of trusted authorities. Rather than replicating the capabilities of the Yellow Pages, SMBmeta's goal is to help small and medium businesses establish a Web presence. However, the two share the approach of annotating Web content with structured information to make it more accessible for various search applications.

#### References

1. E. Amitay et al., "Web-a-Where: Geotagging Web Content," *Proc. 27th Int'l Conf. Research and Development in Information Retrieval*, ACM Press, 2004, pp. 273-280.
2. J. Ding, L. Gravano, and N. Shivakumar, "Computing Geographical Scopes of Web Resources," *Proc. 26th VLDB Conf.*, Morgan Kaufmann, 2000; www1.cs.columbia.edu/~gravano/Papers/2000/vldb00.pdf.
3. K. McCurley, "Geospatial Mapping and Navigation of the Web," *Proc. 10th Int'l Conf. WWW*, ACM Press, 2001, pp. 221-229.
4. R. Bouckaert, "Low-Level Information Extraction: A Bayesian Network-Based Approach," 2002; www-ai.ijs.si/DunjaMladenic/TextML02/papers/Bouckaert.pdf.
5. D. Bricklin, "The SMBmeta Initiative," 2004; www.smbmeta.org.

When Web authors include an address, they make an effort, aided by habit, to include one that is well formed.

contains well-formed postal addresses, landline telephone numbers, or both.

The presence of one or more addresses is a hint about a document's subject that the designers of a search application's relevance ranking algorithms can use as they see fit. One advantage to this approach is that Geosearch is portable. It is a software component that is inserted at a convenient point into a search application's workflow.

## Address recognizers

Geosearch *address recognizers* detect US-conformant addresses consisting of at least a postal code and a preceding state, Canadian postal codes and a preceding province, and North American Numbering Plan (NANP) telephone numbers (US, Canada, Caribbean). Canadian postal codes are particularly well-suited for local search because they have a short but unique format, and, especially in urban areas, they map to accurate latitudes and longitudes.

Geosearch scans all pages for address data. Using brute force to search for US addresses is justified by the fact that such addresses or telephone numbers are found on more than 20 percent of pages.

To internationalize Geosearch, it might be necessary to develop heuristics to determine what types of addresses to look for on a page. Address formats vary by country, and searching for an exhaustive set on each page could be too time-consuming.

Upon finding an address, the address recognizer forwards what it considers the relevant tokens to the geocoder so that it can assign geographic coordinates to the presumed address. Because geocoding is usually expensive compared to scanning, the address recognizer works to reduce the number of false addresses it sends to the geocoder.

## Geosearch observations

For the two years that Geosearch was publicly available, and for the preceding year, Vicinity researchers used these techniques to closely observe Internet-derived local search and identify its strengths, weaknesses, and future opportunities. As a large-scale proof of concept, Geosearch exceeded their expectations.

**Local data permeates the Web.** When Vicinity researchers embarked on developing Geosearch in 1998, they evaluated sets of Web pages provided by several popular search engines and portals. On a consistent basis, more than 20 percent of these pages contained either a well-formed US or Canadian address or NANP telephone number. This percentage remained constant for the two years Geosearch was available on the Internet.

Although Geosearch only looked for North American addresses, the pages it examined were not restricted to North America. Therefore, the percentage of pages with a well-formed address from some country is certain to be higher than the 20 percent that Geosearch found.

**Well-formed addresses are the rule, not the exception.** The efficiency of the address recognition process was a concern to all of the engineering groups the Vicinity researchers worked with. By requiring a well-formed address, the researchers eliminated fruitless work examining text around tokens that marginally looked like part of an address but were not, such as "Chicago-style pizza." It turns out that requiring the combination of a state and postal code is not much of a sacrifice.

In most cases, addresses on Web pages conform to the postal standards used for the delivery of land mail. Occasionally, a postal code that a group of addresses shared was factored out of individual addresses and placed at the start of a table. Overwhelmingly, however, when Web authors include an address, they make the effort, aided by habit, to include one that is well formed. Thus, by promulgating addressing standards for the efficient delivery of land mail, national postal services have made a major contribution toward geoenabling the Web.

If telephone numbers are excluded, Geosearch found at least one address on 15 percent of pages. Some enterprises use a telephone number as the primary contact point. Plumbers, for example, serve a geographic area, and they rely on a phone number rather than a storefront to establish a local presence. A counter example is that customer service phone numbers are probably not interesting for local search. Nationwide customer support numbers, however, are often toll free, and Geosearch did not consider these.

Sometimes the address recognizer found a telephone number, but not an accompanying address that was in fact available. In these cases, the presence of a phone number could trigger more intensive scanning of the surrounding text for an address. The basic results were so encouraging, however, that we did not consider additional work on the address recognizer to be a high priority.

**Addresses are keys to rich exogenous content.** For most people, an address provides enough information to build a mental image of a location in a familiar neighborhood or to use as an index for finding the location on a map. An address is not directly usable for the distance computations and the mapping and routing applications that location-based computing requires. This is the job of geocoding applications that associate an address with a physical location on Earth's surface.

The databases that these applications use represent significant intellectual capital. For example, the US street network product from Geographic Data Technology, a leading provider of geocoding databases, contains more than 14 million addressed street segments, postal and census boundaries, landmarks, and water features. The company processes more than one million changes for this database each month (www.geographic.com/home/productsandservices.cfm).

An address is the key that associates this rich vein of exogenous information with Web content. Addresses are proportionally more valuable for local search because they are computationally easy to detect.

**Addresses are metadata.** The WWW Consortium defines metadata as machine-understandable information for the Web (www.w3.org/metadata). To date, attempts to incorporate metadata into search engine relevancy metrics have not gone well. HTML metatags are ignored because they are either misused or used fraudulently, and metadata standards have no value if they are disregarded. It's interesting to envision the semantic Web that metadata enables, but it's not yet ready for prime time.

These concerns are not persuasive for local search. Geosearch works because of the anomalous but fortunate circumstance that the metadata it depends on is already pervasive on the Internet. An address *is* metadata; its definition predates the Web, but its structure is portable to it.

**Pages with addresses tend to be good quality.** Organizations that put postal addresses on Web pages see the Internet not as a frivolity, but as a way to convey information. Even when a page with an address is unappealing, a quick glance at the site usually leads a user to conclude that the authors don't know how to create a good Web presence, not that they are swindlers or kids with too much time on their hands.

**Local search is about more than commerce.** Internet content reflects what people do—and they do more than shop: They have hobbies, seek like-minded individuals, look for support in times of stress. Sometimes when people do shop, either from preference or necessity, they are not looking for the closest chain store. They are looking for the practitioner of a local craft—a scrimshaw artist in Nova Scotia—or for some activity that is not quite mainstream—worm composting or home solar power generation.

The individual constituencies for the activities people pursue on a daily basis might be small, but taken together they comprise much of the regional information people search for. Some of the most satisfying Geosearch results were for idiosyncratic local content: breast cancer support groups, bird sightings, first-edition rare books, maple syrup (in Vermont), Washington Irving (in Tarrytown, New York).

**The data that characterizes the Internet Yellow Pages is broad, uniform, shallow, and slow to change.**

One hundred years ago the Sears catalog was an innovation for distributing information about mainstream consumer goods. Improvements since then have been around the edges. The overlooked promise of local search is that it makes niche information not routinely found in mail-order or Internet catalogs, the Yellow Pages, or on television or radio, easy to come by. In this it is unrivaled by previous distribution mediums.

## OFFLINE-DERIVED LOCAL CONTENT

In contrast to Internet-derived local content, the data that characterizes the Internet Yellow Pages is broad, uniform, shallow, and slow to change. This data wends a circuitous route from initial compilation to its final destination in IYP listings.

List compilation vendors, whose traditional customers use their products for business mailing lists, sales lead generation, and other direct mail and telemarketing applications, furnish IYP data. The compilers' main data sources are printed telephone directories, which are converted to digital information with OCR devices.

InfoUSA, a leading provider of premium lists, compiles its US list of 14 million businesses from 5,200 phone directories (www.infousa.com). The company augments this phone book data with secondary data sources such as annual reports, SEC filings, business registrations, and postal service change-of-address files. The compilers verify the information they gather by calling businesses. InfoUSA makes 17 million such telephone calls annually.

The IYP is slow to incorporate new and changed information, a shortcoming that is inherent in the source of its data, since telephone books are published annually. List vendors do generate periodic update files, but these updates are not free, and the effort required to merge them into the IYP is not trivial.

More fundamentally, staying current is an elusive goal for decentralized information that is compiled centrally. Telephone directories are out of date even at the moment of publication. Then, list vendors must correlate changes from their incoming data streams—the 5,200 directories, telephone verification calls, change-of-address files, and so forth. Each periodic update includes only a fraction of the changes in a vendor's possession, and it includes no changes that have occurred but are still in the pipeline.

The IDYP is a directory of local businesses, similar to the IYP but richer in content.

Another problem with using compiled lists as a source for the IYP is that the consumer is not its primary market. The lists are flat structures without sufficient expressive power to convey the hierarchical and variable structure of many enterprises, specifically those with multiple external points of contact.

This missing information corresponds to some of the most dog-eared pages in printed directories: individual departments and physicians in hospitals and medical practices, group practices of all sorts, and municipal information. Even if this deficiency were somehow fixed, IYP service providers would still need to reflect the richer structure in the online databases they build from the compiled lists.

For all their shortcomings, the compiled lists from which the IYP is derived are authoritative and trusted sources of business information—characteristics that are not duplicated elsewhere. The clerks making those calls provide real value. Even if the information in the IYP already exists on the Internet, or will sometime soon, it is in a chaotic form, and there is no repeatedly reliable way to access it. The value of the compiled lists is data aggregation, an area in which local search has not yet contributed.

## DECENTRALIZED DATA GATHERING: THE INTERNET-DERIVED YELLOW PAGES

The central challenge for local search is to move the job of aggregating and verifying local information closer to the sources of knowledge about that information. The human and electronic knowledge about local information is decentralized—geographically localized—and the Internet is a decentralized medium. Having decentralized tools for gathering this data is desirable as well.

### Trusted authorities

The IDYP is a directory of local businesses, similar to the IYP but richer in content. The essential difference is that the information in the IDYP is derived directly from the Internet, not from offline sources.

The IDYP's viability depends on intermediaries, trusted authorities who can vouch for the information the IDYP provides and can perform the role of content aggregator for entities without a direct Web presence. Organizations that have relationships of trust with both the public and the entities whose information they are certifying or creating can perform this gatekeeper role.

Two examples of organizations that can serve as gatekeepers are those based on geography, such as a chamber of commerce, and those based on market segment, such as a trade organization. A primary function of both types of representative organizations is to collate and disseminate accurate information on behalf of their members. Both types of organizations are often conversant with Web technology, and they can function as proxies for constituents who don't have their own Web presence. While there are 14 million businesses in the US—most of them small—chambers of commerce and trade organizations number in the thousands.

Preventing fraudulent interlopers from compromising the integrity of their constituents' information is also in the best interests of these gatekeepers. For chambers, this is conceptually and practically as simple as ensuring that each member it verifies or submits to the IDYP does indeed have a shop on Main Street.

The first function of a trusted authority is either to submit information to the IDYP on behalf of a member or to certify information a member has directly submitted to the IDYP. The second function, policing, is aimed at minimizing the amount of fraudulent or misleading data that makes its way into the IDYP.

**Proxy mode.** In proxy mode, trusted authorities are intermediaries for members who want a presence in the IDYP but do not want to interact directly with the Internet.

For example, a licensed hotdog-stand vendor with no interest in using the Internet would work with a representative at the chamber of commerce to get the right information into the IDYP. A hypothetical entry for this vendor would indicate that the stand is open from two hours before an event until one hour afterwards, provide the stand's location in the stadium, and state the type and price of the hotdogs, drinks, and condiments he sells. If, at the last minute, the vendor finds he will not be at an event, he can ask his contact to update his IDYP entry. This example is contrived, of course—but try finding information on street vendors in the Yellow Pages.

**Authenticate mode.** A trusted authority uses a Web interface to help create the structured information the IDYP requires for the members under its purview. The trusted authority releases this information to an IDYP server, at which point it becomes generally available. An entity can directly submit information to an IDYP server as long as the submittal refers to at least one valid registration with at least one trusted authority.

### Policing

Unlike a purely virtual search, the subject matter in local search has a physical existence that can be confirmed. Therefore, local search is more resistant to fraud than are purely virtual searches. In the IDYP model, if no trusted authorities vouch for a business, it will not be included in the IDYP. Still, we must assume that efforts will be made to game the system and that some businesses will be tempted to misrepresent themselves.

The Internet's potential to provide assurances about local enterprises exceeds that of current directory services. It isn't possible to rely on the Yellow Pages to provide guidance about a business's reputation. The IDYP, however, can augment its information with various data sources, such as Better Business Bureaus, independent reviews, and public data. In addition, the IDYP can use practices that have become popular on the Internet for rating products, services, and sellers.

## IDYP OPERATION MODES

Geosearch found that at least 20 percent of Web pages include an overt geographic qualifier. Even if every local enterprise eventually registers with a trusted authority, the Web will still contain much local content that is not known to the IDYP.

Geosearch's strength is that it finds local content in place, without requiring Web authors to change their routines for publishing that content. To integrate its content with local content on the Web that is not part of the IDYP, the IDYP supports two modes of operation. In one mode it supplies local search metadata to authorized applications; in the other, it is a stand-alone directory application.

### Local search metadata

In the local search metadata mode, the IDYP makes its content available to subscribing applications. Subscribers are bound to use IDYP data in conformance with the policies and standards the IDYP sets forth. Trusted authorities and individual businesses can also specify directives on how subscribers use their information.

As a part of the page indexing process, a subscribing search application seeks associated IDYP information for the page it is indexing. If the page is authorized for local search, the search application includes some portion of the IDYP metadata in the index it builds for the page. In this way, IDYP data is incorporated into the general Web corpus.

**In the IDYP model, if no trusted authorities vouch for a business, it will not be included in the IDYP.**

A URL provides the connection between IDYP data and data on the Web. When an enterprise or its trusted authority creates its IDYP entry, it specifies a Web page address with which the IDYP entry is associated. This is the page to which the search application adds the IDYP metadata.

For a member who doesn't have a direct Web presence, the trusted authority creates one or more pages that contain formatted content derived from the member's IDYP entry. The trusted authority either establishes a domain for the member or guarantees that the pages it creates for the member have persistent URLs.

A trusted authority might choose to generate pages for all its members. This would allow it to establish a consistent look and feel for its constituents. IDYP pages generated for members that already have established Web sites would contain links back to these existing pages.

The IDYP provides an imprimatur for pages that are relevant for local search. To accommodate pages with local content unknown to the IDYP, search applications can support either strict or nonstrict local searches.

In strict mode, the search application only considers pages that are known to the IDYP. In nonstrict mode, the search application uses its own heuristics for gauging which pages are relevant, and can return a mixed set of pages, some known to the IDYP, some not. If the search application tags the results that are known to the IDYP, users can decide for themselves how important the IDYP imprimatur is. It will be more valuable for Yellow Pages-like searches, less so for idiosyncratic ones.

### Stand-alone local directory

Given the popularity of search engines and portals as user interfaces, observers might anticipate that the IDYP's main role is to provide metadata for these applications. However, as a self-contained local directory, the IDYP can provide powerful features that are difficult to incorporate into a general-purpose search engine. Important, too, is that the IDYP should not depend on any particular search application for its promulgation. Its status as a stand-alone application ensures its independence.

**Standard data representation.** IDYP information is represented in XML. In addition to a standard core of attributes, industry groups can define customized extensions—known as XML schemas. An "hours of operation" attribute, for example, is part of the standard core, since virtually all businesses use it—although today's IYP does not include even this basic information. The XML Schema for restaurants should allow queries about the catch of the day at the local seafood house.

**Rich categorization hierarchy.** The business categorization schemes used by the print and Internet Yellow Pages are cursory. The Internet has spawned much work on commerce-oriented ontologies and user interfaces that are broadly applicable to local search.

**Local search query language.** A rich stratum of metadata will facilitate the construction of a local search query language with more expressive power than the Boolean keyword languages that current search engines use.

The parlance of local search is constrained—a variation of who, what, where, when, and how much: Who provides what service or product? Where is the provider located? When can I see the product? How much does it cost? For example: "Where can I buy stylish children's clothing on sale, within 10 miles of home, open late on Saturday evening?"

**Short update latency.** The time interval between an enterprise making a change and having that change reflected in the IDYP is brief, converging on instantaneous. We can define "the last croissant" heuristic, which states that the IDYP reaches optimal efficiency when an urban bakery can use it successfully to advertise a sale on its remaining bakery items 30 minutes before closing.

Geosearch, a geoenabled search engine that allows people to search for Web pages that contain geographic markers within a specified geographic area, demonstrates that the Internet is a rich source of local content. It also demonstrates the many advantages that postal addresses have as a key for accessing this content, especially when the content pertains to the activities of daily life. Postal addresses are ubiquitous, unambiguous, standardized, computationally easy to detect, and necessary for accessing the rich and precise content of geocoding databases.

The Internet Yellow Pages, currently the main source of local content on the Internet, are reliable, but they are also shallow, slow to change, centralized, and expensive. Their primary data sources are printed telephone directories. They do not use the Internet's resources in any meaningful way.

Local search today provides a poor user experience because it does little more than package old data for a new medium. The richest source of local content can and should be the Internet itself, but marshaling this resource requires developing an infrastructure such as the Internet-Derived Yellow Pages to organize and manage its content.

The IDYP is a structured database that relies on trusted authorities, such as chambers of commerce or trade associations, to certify the information it contains. The IDYP can function either as a standalone directory or as a source of metadata for search applications. A search application uses IDYP metadata to augment the information it maintains for Web pages that have local content. In this way, local search metadata is integrated into the general Web corpus. ■

## References

1. J. Horrigan, Pew Internet Project, "Pew Internet Project Data Memo," Apr. 2004; www.pewinternet.org/pdfs/PIP_Broadband04.DataMemo.pdf.
2. J. Horrigan, L. Rainie, Pew Internet Project, "The Broadband Difference," June 2002; www.pewinternet.org/pdfs/PIP_Broadband_Report.pdf.
3. Kelsey Group & Bizrate.com, "Local Search Now 25% of Internet Commercial Activity," Feb. 2004; www.kelseygroup.com/press/pr040211.htm.
4. S. Kerner, "Majority of US Consumers Research Online, Buy Offline," Oct. 2004; www.clickz.com/stats/markets/retailing/article.php/3418001.
5. G. Sterling, "Is 2004 the Year of Local Search?" Dec. 2003; www.imediaconnection.com/content/2343.asp.

## Acknowledgments

The author thanks the many people involved with making Geosearch happen, including the following—from Vicinity: Jeff Doyle, Charlie Goldensher, Jerry Halstead, Dwight Aspinwall, Dave Goldberg, Faith Alexandre, Darius Martin, Kavita Desai, and Eric Gottesman; and from Northern Light: Marc Krellenstein, Mike Mulligan, Sharon Kass, and Margo Rowley.

***Marty Himmelstein** is a software consultant. His interests include database systems, spatial databases, and Web technologies. He received an MS in computer science from SUNY Binghamton. He is a member of the IEEE Computer Society and the ACM. Contact him at marty@longhill.com.*