Structured Databases on the Web: Observations and Implications*

Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, Zhen Zhang
Computer Science Department
University of Illinois at Urbana-Champaign
{kcchang,binhe,cli,mppatel2,zhang2}@cs.uiuc.edu

ABSTRACT

The Web has been rapidly "deepened" by the prevalence of databases online. With the potentially unlimited information hidden behind their query interfaces, this "deep Web" of searchable databases is clearly an important frontier for data access. This paper surveys this relatively unexplored frontier, measuring characteristics pertinent to both exploring and integrating structured Web sources. On one hand, our "macro" study surveys the deep Web at large, in April 2004, adopting the random IP-sampling approach, with one million samples. (How large is the deep Web? How is it covered by current directory services?) On the other hand, our "micro" study surveys source-specific characteristics over 441 sources in eight representative domains, in December 2002. (How "hidden" are deep-Web sources? How do search engines cover their data? How complex and expressive are query forms?) We report our observations and publish the resulting datasets to the research community. We conclude with several implications (of our own) which, while necessarily subjective, might help shape research directions and solutions.

1. INTRODUCTION

In the recent years, the Web has been rapidly "deepened" by the massive networked databases on the Internet: While the *surface Web* has linked billions of static HTML pages, it is believed that a far more significant amount of information is "hidden" in the *deep Web*, behind the query forms of *searchable* databases. Using overlap analysis between pairs of search engines, a July-2000 white paper [1] estimated at least 43,000-96,000[1] "deep Web sites," and claimed 550 billion hidden pages in the deep Web, or 550 times larger than the surface Web.

These databases are often also referred to as the *hidden* or *invisible* Web: The perception naturally arises: Since such information cannot be accessed directly through static URL links, they are only available as responses to dynamic queries submitted through the *query interface* of a database. Because current crawlers cannot effectively query databases, such data are invisible to traditional search engines, and thus remain largely hidden from users.

This paper surveys databases on the Web, for characteristics pertinent to their exploration and integration. The survey is based on our experiments in April 2004 for the deep Web at large (Section 3) and December 2002 for source-specific characteristics (Section 4). With its massive sources, this deep Web is an important yet largely-unexplored frontier for data integration. While our research community has actively studied integration techniques, such *large-scale integration* is not a traditional focus. We hope a reality check will help identify the challenges and sketch the landscape, for motivating and guiding our future research efforts.

Specifically, our survey focuses on *structured* databases on the Web, which return *structured objects* with attribute-value pairs (*e.g.*, a `Book` source like *amazon.com* returns books with author, title, *etc.*). Thus, our focus essentially distinguishes *unstructured* databases, which provide data objects as unstructured media (*e.g.*, texts, images, audio, and video). We believe such distinction is both desired and necessary: First, such structured or "relational" data are traditionally of greater interest to the database community. Second, structured sources necessarily imply different paradigms and techniques from unstructured sources.

We design our survey to center around the essential tasks for effectively accessing the deep Web: That is, while there are myriad useful databases, how can a user *find* the right sources and *query* them in a right way? Consider user Amy, who just joined a university as a new professor. To find a house, where can she look for real estate listings in her town? (*Realtor.com*.) Where can she study for buying a new car? (*Cars.com*, *Edmunds.com*.) To plan her research agenda, how can she find emerging topics in databases? (Try *DBLP* Bibliography Search.) After surviving source hunting, Amy will realize that she has to learn the grueling details of querying, which can be a major headache especially when there are multiple sources.

Our survey thus studies issues related to these dual essential tasks: First, for *exploration* (*i.e.*, to help Amy find sources), our *macro* study surveys the deep Web at large: What is its scale? How many databases are there? Where to find "entrance" to them? How many are structured databases? What is the coverage of deep-Web directories? What is the category distribution of sources? Second, for *integration* (*i.e.*, to help Amy query sources), our *micro* study surveys source characteristics: How "hidden" are Web sources? How do search engines cover their data? How complex are their query interfaces? How complex are their queries?

To our knowledge, this survey is the first "open-source," fully documented study of the deep Web, with a specific focus on *structured* databases, for both the macro and micro perspectives. Most Web scale characterization efforts have focused on the surface Web, *e.g.*, [2]. The pioneering study [1] of the deep Web has since opened wide interests in this area; however, in comparison, it differs in several aspects: 1) It studies Web "search sites" in a seemingly broader sense, without giving explicit qualification of such sites. 2) It uses proprietary methods, which result in much unexplained detail. 3) It studies mainly about the "macro" but lacks the "micro" perspective.

*This material is based upon work partially supported by NSF Grants IIS-0133199 and IIS-0313260. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

[1]The white paper [1] then claimed 200,000 sites to account for "under count due to lack of randomness."

domain	*sources*	*domain*	*sources*
Airfares	50	Hotels	39
Automobiles	97	Jobs	52
Books	59	Movies	69
CarRentals	24	MusicRecords	51

Figure 1: Domain-specific dataset: 441 sources in 8 domains.

Finally, based on our findings, we suggest several likely implications. While our interpretation of the results and our conjectures are necessarily subjective, we believe they are at least well motivated by the survey, and are likely to shed insights for our future research. Our main conclusions are– 1) in terms of *problems*: large-scale integration is a real challenge, which likely will mandate dynamic and ad-hoc integration requirements; and 2) in terms of *solutions*: holistic-integration approaches, which discover integration semantics by globally exploiting shallow clues across many sources, are likely to be a key technique for enabling large-scale integration.

We start in Section 2 by discussing our experimental setup and methodologies. Section 3 reports the results of our macro study, and Section 4 our micro study. We then discuss our implications in Section 5. Finally, Section 6 reviews the related work and Section 7 concludes the paper.

2. EXPERIMENTAL SETUP

Our survey intended to study both the "macro" characteristics of the deep Web at large and the "micro" characteristics of sources in some representative domains. We thus configured two groups of experiments, each with different datasets. First, we adopted the random IP-sampling approach to acquire Web sites from a sample of 1 million randomly-generated IP (Internet Protocol) addresses. These sampled sources constitute the dataset for our macro survey. Second, for our micro study, we manually collected 441 sources in 8 representative domains.

2.1 Randomly-Sampled Dataset

We performed our "macro" experiments in April 2004 to study the deep Web at large: its scale in particular. There are mainly two approaches for such Web size characterization. The first scheme, *overlap analysis*, estimates the Web size by extrapolating from the overlap size between two independently and randomly sampled collections, *e.g.*, search engines. Such estimates, as [1, 2] show, tend to result in great inconsistencies when different search engines are used, because the independent and random sampling assumptions may not hold. We thus adopt the *random IP-sampling* method, which estimates the Web size by testing randomly-sampled IP addresses. This scheme assumes that Web servers are uniformly distributed in the entire IP space. The assumption seems more realistic, and the results are in fact more consistent and stable.

Our experiment sampled 1,000,000 unique randomly-generated IPs, from the 2,230,124,544 valid IP addresses (the entire space after removing the reserved and unused IP ranges according to [3]). For each IP, we used an HTTP client, the GNU free software wget [4], to make an HTTP connection to it and download HTML pages. The results show that among these 1,000,000 IPs, 2256 IPs have publicly accessible Web sites, by responding to our HTTP requests. These sources constitute our sample of the Web, based on which we further examined the presence of Web databases. Section 3 reports our survey on this sampled dataset.

2.2 Domain-Specific Dataset

We performed our "micro" experiments in December 2002 to study *per-source* characteristics of deep Web sources. To inspect

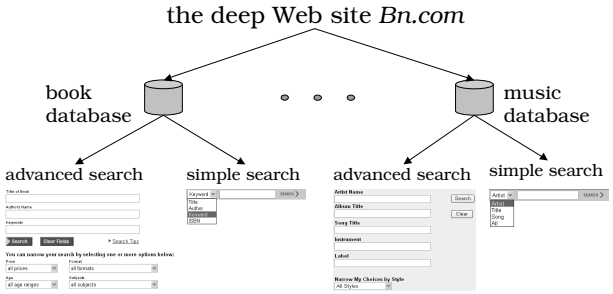


Figure 2: Site, database, and interface.

any potential domain-specific implications, we took a *domain-centered* approach, in which we studied sources in several representative domains. We manually collected deep web sources using Web directories (*e.g.*, InvisibleWeb.com, BrightPlanet.com, Web-File.com) and search engines (*e.g.*, Google.com). In particular, we collected 441 sources in eight domains: Airfares, Automobiles, Books, Car Rentals, Hotels, Jobs, Movies and Music Records. Figure 1 summarizes our dataset. We have released this dataset, as part of the *UIUC Web Integration Repository* [5], available online at *http://metaquerier.cs.uiuc.edu/repository*. In particular, the dataset gives the complete list of sources we studied in this survey.

3. THE MACRO: DEEP WEB AT LARGE

This section presents our macro survey of the deep Web at large. Our focus is centered around the challenge of *exploring* databases on the Web, *i.e.*, finding where they are (as Section 1 introduced). Our survey thus intended to address both the *scale* of the deep Web, and the *coverage* of current directory services, with an emphasis on structured databases (although we also measured unstructured sources). For this set of experiments, we adopted the IP-sampling approach (Section 2.1).

In our survey, we distinguished three related notions for accessing the deep Web– site, database, and interface: A *deep-Web site* is a Web server that provides information maintained in one or more back-end *Web databases*, each of which is searchable through one or more HTML forms as its *query interfaces*. For instance, as Figure 2 shows, *bn.com* is a deep-Web site, providing several Web databases (*e.g.*, a book database, a music database, among others) accessed via multiple query interfaces (*e.g.*, "simple search" and "advanced search"). Note that, our survey considered only unique interfaces and removed duplicates– Many Web pages contain the same query interfaces repeatedly, *e.g.*, in *bn.com*, the simple book search in Figure 2 is present in almost all pages.

As our survey specifically focuses on online databases, we differentiated and excluded non-query HTML forms (which do not access back-end databases) from query interfaces. In particular, HTML forms for login, subscription, registration, polling, and message posting are not query interfaces. Similarly, we also excluded "site search," which many Web sites now provide for searching HTML pages on their sites– These pages are statically linked at the "surface" of the sites; they are not dynamically assembled from an underlying database.

(Q1) *Where to find "entrances" to databases?* To access a Web database, we must first find its *entrances*– *i.e.*, query interfaces. How does an interface (if any) locate in a site– *i.e.*, at which depths? For each query interface, we measured the *depth* as the minimum number of hops from the root page of the site to the interface page. For this study, as it required deep crawling of Web sites, we analyzed $\frac{1}{10}$ of our total IP samples, *i.e.*, a subset of 100,000 IPs. We tested each IP, by making HTTP connections, and found 281 Web servers. Exhaustively crawling these servers to depth 10,

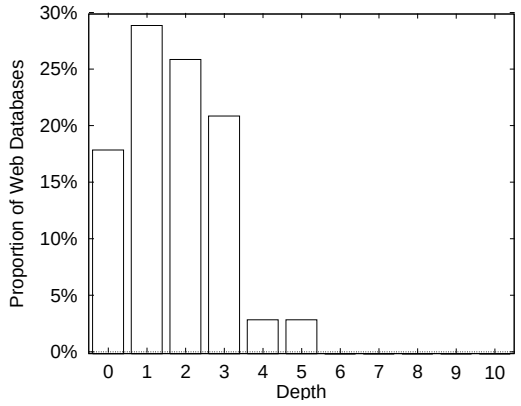


Figure 3: Distribution of Web databases over depth.

	Sampling Results	*Total Estimate*
Deep Web sites	126	307,000
Web databases	190	450,000
– unstructured	*43*	*102,000*
– structured	*147*	*348,000*
Query interfaces	406	1,258,000

Figure 4: Sampling and estimation of the deep-Web scale.

we found 24 of them are deep Web sites, which contained a total of 129 query interfaces representing 34 Web databases.

We found that query interfaces tend to locate shallowly in their sites– None of the 129 query interfaces had depth deeper than 5. To begin with, 72% (93 out of 129) interfaces were found within depth 3. Further, since a Web database may be accessed through multiple interfaces, we measured its depth as the minimum depths of all its interfaces: 94% (*i.e.*, 32 out of 34) Web databases appeared within depth 3; Figure 3 reports the depth distribution of the 34 Web databases. Finally, 91.6% (*i.e.*, 22 out of 24) deep Web sites had their databases within depth 3. (We refer to these ratios as *depth-three coverage*, which we will guide our further larger-scale crawling in **Q2**.)

(Q2) *What is the scale of the deep Web?* We then tested and analyzed all of the 1,000,000 IP samples to estimate the scale of the deep Web. As just identified, with the high depth-three coverage, almost all Web databases can be identified within depth 3– We thus crawled to depth 3 for these 1 million IPs.

The crawling found 2256 Web servers, among which we identified 126 deep Web sites, which contained a total of 406 query interfaces representing 190 Web databases. Extrapolating from the $s = 1,000,000$ unique IP samples to the entire IP space of $t = 2,230,124,544$ IPs, and accounting for the depth-three coverage, we estimate the number of deep Web sites as $126 \times \frac{t}{s} \div 91.6\% = 307,000$, the number of Web databases as $190 \times \frac{t}{s} \div 94\% = 450,000$, and the number of query interfaces as $406 \times \frac{t}{s} \div 72\% = 1,258,000$ (the results are rounded to 1000). Table 4 summarizes the sampling and the estimation results. By their ratios, we also observed the "multiplicity" of access on the deep Web. In average, each deep Web site provides 1.5 databases, and each database supports 2.8 query interfaces.

The earlier survey of [1] estimated 43,000 to 96,000 deep Web sites by overlap analysis between pairs of search engines. Although the white paper has not explicitly qualified what it measured as a "search site," by comparison, it is still evident that the scale of the deep Web is well on the order of 10^5 sites. Further, it has been expanding, resulting in 3-7 times increase in 4 years (2000-2004).

	Number of Web Databases	*Coverage*
completeplanet.com	70,000	15.6%
lii.org	14,000	3.1%
turbo10.com	2,300	0.5%
invisible-web.net	1,000	0.2%

Figure 5: Coverage of deep-Web directories.

(Q3) *How "structured" is the deep Web?* While information on the surface Web is mostly unstructured HTML text (and images), how is the nature of the deep-Web data different? We classified Web databases into two types: 1) *unstructured* databases, which provide data objects as unstructured media (*e.g.*, texts, images, audio, and video), and 2) *structured* databases, which provide data objects as structured "relational" records with attribute-value pairs. For instance, *cnn.com* has an unstructured database of news articles, while *amazon.com* has a structured database for books, which returns book records (*e.g.*, title = "gone with the wind", format = "paperback", price = $7.99).

By manual querying and inspection of the 190 Web databases sampled, we found 43 unstructured and 147 structured. We similarly estimate their total numbers to be 102,000 and 348,000 respectively, as Table 4 also summarizes. Thus, the deep Web features mostly structured data sources– with a dominating ratio of 3.4:1 versus unstructured sources.

(Q4) *What is the coverage of deep-Web directories?* Besides traditional search engines, several deep-Web portal services have emerged online, providing deep-Web directories which classify Web databases in some taxonomies. To measure their coverage, we surveyed four popular deep-Web directories, as Figure 5 summarizes. For each directory service, we recorded the number of databases it claimed to have indexed (on their Web sites). As a result, *completeplanet.com* was the largest such directory, with over 70,000 databases[2]. As Figure 5 reports, compared to our estimate, it covers only 15.6% of the total 450,000 Web databases. However, other directories covered even less, in the mere range of $0.2\% - 3.1\%$. We believe this extremely low coverage suggests that, with their apparently manual classification of Web databases, such directory-based indexing services can hardly scale for the deep Web.

(Q5) *What is the subject distribution of Web databases?* With respect to the top-level categories of the *yahoo.com* directory as our "taxonomy," we manually categorized the sampled 190 Web databases. Figure 6 shows the distribution of the 14 categories: Business & Economy (be), Computers & Internet (ci), News & Media (nm), Entertainment (en), Recreation & Sports (rs), Health (he), Government (go), Regional (rg), Society & Culture (sc), Education (ed), Arts & Humanities (ah), Science (si), Reference (re), and Others (ot). The distribution indicates great subject diversity among Web databases, suggesting that the emergence and proliferation of Web databases are spanning well across all subject domains.

4. THE MICRO: DOMAIN STUDIES

Beyond our macro study, we also investigated 441 sources (Section 2.2) to survey per-source characteristics. These sources were from 8 representative domains– Our study also intended to identify, if any, domain-specific implications. For this set of experiments, we focus on the challenge of *integrating* databases on the Web, *i.e.*, accessing and querying them (as Section 1 introduced).

We performed two groups of experiments. First, in [Q6–7] we

[2]However, we noticed that *completeplanet.com* also indexed "site search," which we have excluded; thus, its coverage could be overestimated.

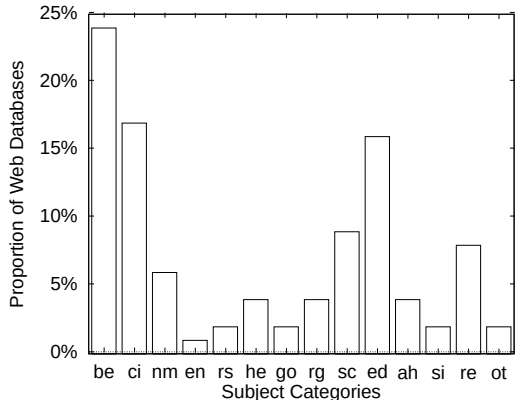


Figure 6: Distribution of Web databases over subject category.

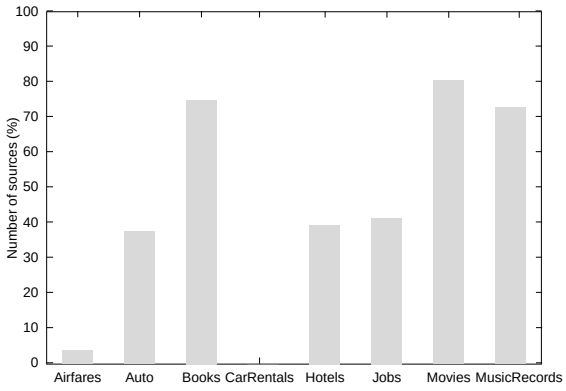


Figure 7: Hiddenness: Ratio of sources with browse interfaces.

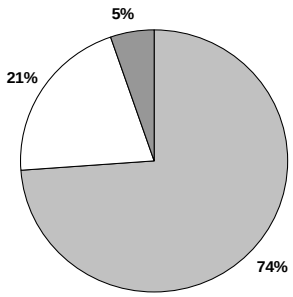
studied the *accessibility* of sources: How hidden are their data? How search engines have crawled their data? Second, in [Q8–11], we studied the various aspects about querying sources – by investigating the "complexity" of their query interfaces. We believe for any attempt to integrate structured databases (*e.g.*, query mediation), it is essential to cope with these query interfaces, since data must be retrieved with queries.

(Q6) *How "hidden" are data on the deep Web?*

The deep Web is often referred to as the "hidden" or "invisible" Web. The impression has naturally arisen from that data can "only" be accessed through query interfaces, and thus are hidden from any typical crawlers that follow hyperlinks. This "query-only" access mode essentially distinguishes databases on the Web (the deep Web) from the rest of the link-based contents (the surface Web). To verify the restrictions as well as alternatives, we validated whether the deep Web is mostly hidden.

We examined, for each source, whether there are *navigational* paths to reach its data, which essentially "surface" the data. Such navigational paths are typically provided by a *browse interface* for accessing data by navigating some data-classification hierarchy (*e.g.*, *Amazon.com* allows browsing books in a subject hierarchy). Thus we measured the *hiddenness* of the deep Web sources by checking the availability of browse interfaces. Such navigational access, when available, provides link-based access to data. (However, we did not further measure if such navigational access indeed reaches all the data that the corresponding query-based access does.) Figure 7 reports, for each domain, the ratio of such "open" sources.

The results seem somewhat surprising: The deep Web is not "entirely" hidden; for some domains, there often exists navigational access to data. That is, the hiddenness varies across domains:



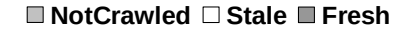


Figure 8: Coverage and freshness of Google cache.

While some domains (*e.g.*, Airfares, CarRentals) usually do not support browse interfaces, others (*e.g.*, Books, Movies) tend to be quite "open." Such variation might have resulted naturally from the "dynamism" of data. For instance, Airline reservation data are highly dynamic and seasonal, thus making them harder to maintain in static links than other relatively more static data (*e.g.*, Books). Another possible reason is the "browsability" of data, *i.e.*, whether there exist some natural, commonly-accepted organizational hierarchies to browse the data (*e.g.*, Books, Movies).

(Q7) *How do search engines cover deep-Web sources?* As a consequence of [Q6], since deep-Web sources may not be hidden, is it possible to "crawl-and-warehouse"as search engines do for the surface Web? To answer this question, we investigated how a typical search engine "warehouses" such data, for both coverage and freshness. In particular, we use Google (`google.com`) because it supports access to the "cached" pages.

We randomly chose 10 sources in each domain. For each source, first, we manually selected some objects (result pages) as test data (without any particular bias) by querying the data source (*e.g.*, *Amazon.com*). We then, for each object collected, used Google's "advanced search" to check if Google crawled its page and if the page contained up-to-date information. We formulated a query and submitted to Google to match the test object. (For instance, we used distinctive phrases occurred in the object page as keywords and limited the search to only the source site.) For a cached page, we further checked if it was fresh, by comparing its information to the source object (*e.g.*, the price may change).

Figure 8 reports the distribution of objects that were not-crawled, crawled-but-stale, and fresh. First, most deep-Web data are simply not covered by Google. Second, even covered, most cached data are stale. The freshness is only 5%. Thus, the "crawl-and-warehouse" approach might not work well for deep-Web data.

(Q8) *How large is query-interface schema?* Each query interface supports queries on some *attributes* (*e.g.*, title for Books); these attributes form the "schema" aspect of query interfaces. As our survey focuses on structured databases, such schema information is essential. We thus measured the number of attributes, as the *schema size*, for each source query interface.

Figure 9 shows the distribution of schema sizes across all domains (individually and overall). For instance, consider the most frequent sizes: for Jobs, 25% sources had 5 attributes, and for Car Rentals 28% had 8. What's the overall most-frequent size? 18% interfaces had size 4.

Figure 10 shows the smallest, largest, and average schema sizes for each domain and overall. First, some domains tend to be more

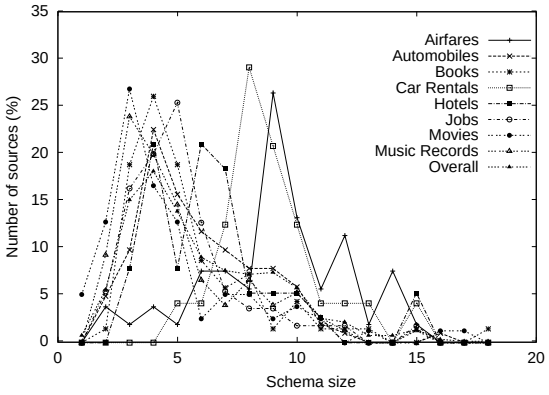


Figure 9: Distribution of schema sizes.

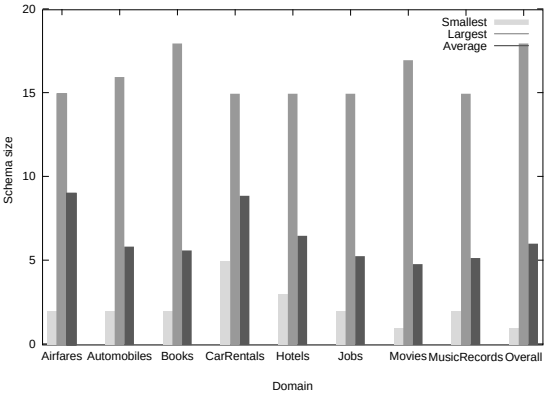


Figure 10: Smallest, largest and average schema sizes.

complex: Airfares and Car Rentals have an average schema size of 9, which is larger than the overall average. Second, some domains show a more significant difference across sources (*i.e.*, larger difference between the smallest and largest), such as Books and Music Records; others are more uniform, such as Car Rentals. Overall, across all sources, the smallest size of schema is 1, the largest 18, and the average 6.

(Q9) *How complex are the "schema vocabularies" for query interfaces?* Consider attributes for querying as the *schema vocabulary* for query interfaces– Do sources in the same domain somehow share a schema vocabulary? How complex is such an aggregate vocabulary? Here we report our analysis of our sources for this schema complexity.

For attribute comparison, as preprocessing, we applied several simple and common normalization steps to identify the same attributes with slightly different textual appearances. Each attribute is normalized by three steps: stopwords removal (*e.g.*, "the", "of"), stemming (authors becomes author), and alphabetical ordering (book titles and title of books both become book title).

First, we see the *clustering* behavior among the schema attributes. An attribute tends to relate to certain others, and they together form a *locality* of co-occurring attributes (*e.g.*, author tends to cluster with title, and make with model). Further, these natural localities quite precisely correspond to the structural *domains* of their sources (*e.g.*, Books, Automobiles). Figure 11 plots how attributes (the y-axis) occur in sources (the x-axis), so that a dot at (x, y) indicates that the schema of source x contains attribute y. Note that sources are ordered according to their domains, and attributes according to their order of first-occurrence along these ordered sources. Observe that each densely-dotted triangle along the diagonal represents an attribute locality, which is

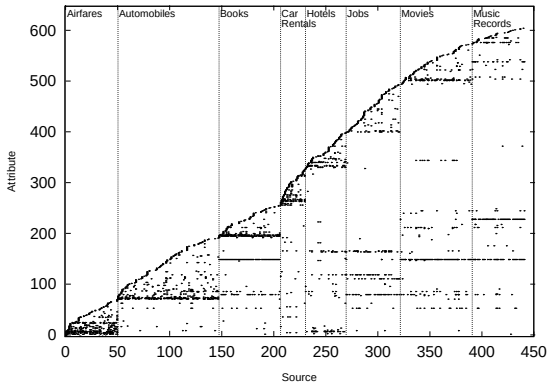


Figure 11: Attribute distributions over source domains.

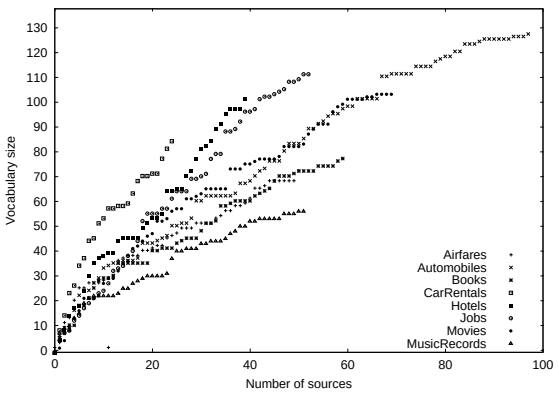
also squarely aligned with the domain boundaries of sources on the x-axis.

Second, we observe the *convergence* behavior: The aggregate schema vocabulary of sources in the same domain tends to converge at a relatively small size. Figure 12(a) analyzes the growth of vocabularies as sources increase in numbers for each domain. The curves indicate the convergence of vocabularies– Since the vocabulary growth rates (*i.e.*, the slopes of these curves) decrease, as sources proliferate, their vocabularies will tend to stabilize. For instance, for Automobiles, 80% (103/129) attributes are observed at 63^{th} sources out of 97 sources. Such convergence effects will be more obvious, if we weight the vocabulary growth by the "importance" of a new attribute– For the purpose of integration, an attribute that occurs in many sources will be more important. We thus further analyze the growth of *frequency-weighted* vocabulary size for each domain, as shown in Figure 12(b). To quantify, let the *frequency* of an attribute be the number of sources in which it occurs. When counting the vocabulary size, each attribute is now weighted by its frequency in the corresponding domain. We see a very rapid convergence– In other words, as we see more sources, the addition of attributes tends to be rather insignificant.

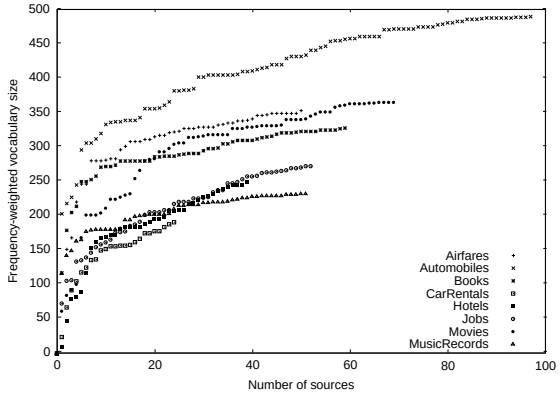
Third, we see extremely non-uniform Zipf-like distributions of attribute frequencies. (Thus some attributes are much more "significant" than others.) Figure 13 orders the frequencies of attributes over their ranks. It is interesting but perhaps not surprising to observe that the distribution obeys the Zipf's law [6]: The frequencies are inversely proportional to their ranks. Many low-ranked attributes thus rarely occur; in fact, 61% (368/607) attributes occur in only one source. Further, frequent attributes dominate: we observe that the top-20 attributes, or 3.3% (20/607) attributes, constitute 38.4% (953/2484) of all the occurrences. What are the most "popular" attributes across all these sources? The top 5 frequent attributes are, in this order, title, keyword, price, make, and artist.

Finally, we see the *linking* behavior. As shown in Figure 11 the attributes from different domains are not isolated, but related– as manifested by the horizontal dotted lines below the diagonal triangles (*i.e.*, outside the localities), which span across several domains. Such "linkages" indicate natural semantics connections between different domains, reflected by their common attributes. Further, the linkages capture the natural "proximity" of domains very well– *i.e.*, some domains are more related than others. For instance, Movies and Music Records are heavily linked (by several horizontal lines), which indicates their intrinsic proximity. Similarly, Airfares, Hotels, and Car Rentals form another related "cluster."

Overall, the findings seem to shed light in coping with the myr-



(a) Vocabulary growth over sources in each domain.



(b) Frequency-weighted vocabulary growth.

Figure 12: Convergence of schema vocabularies.

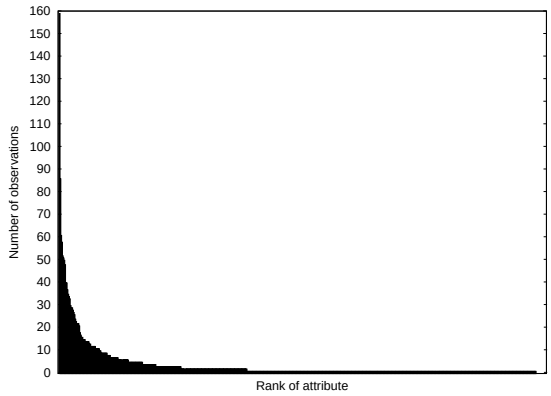


Figure 13: Frequencies over ranks for all attributes.

iad sources on the deep Web– by leveraging their potential regularities (as Section 5 will discuss). We have observed that, while sources proliferate, their aggregate "complexity" does not grow indefinitely, but instead demonstrates certain statistical regularities. In particular, their schema "vocabularies" tend to cluster, converge, and interlink, revealing some hidden structures. The Zipf-distribution also hints an effective strategy using the classic 80-20 rule– that a few perhaps dominate all.

(Q10) *How complex are the "constraint vocabularies" for queries?* We have just seen in [Q9] the regularities of schema vocabularies– To what extent can we see such concerted complexity on the deep Web? To further validate, we also analyzed the "building blocks" for query interfaces– *i.e.*, the *constraint patterns* that express atomic conditions in query forms. For example, as Figure 14 shows, the query interface of *barnesandnoble.com* has seven constraints (*e.g.*, on title and price, *etc.*) and *autos.msn.com* five (*e.g.*, on make, model, *etc.*). We similarly refer to these constraints as the *constraint vocabulary* for queries. (Any query language, such as SQL, has such a vocabulary for formulating queries, *e.g.*, [age $>$ 18].) Observe that these constraints can be abstracted to share some common syntactic patterns of expression: For instance, the format constraint of *barnesandnoble.com*, the category of *autos.msn.com*, among others, all share the pattern of [*attribute equal enumeration*].

This observation again reveals that the query vocabulary of online sources might not be entirely chaotic. What is this vocabulary? How large? For these questions, we manually surveyed 3 domains (from our dataset): Books, Automobiles, and Airfares. We chose these domains because they are schematically dissimilar and semantically unrelated.

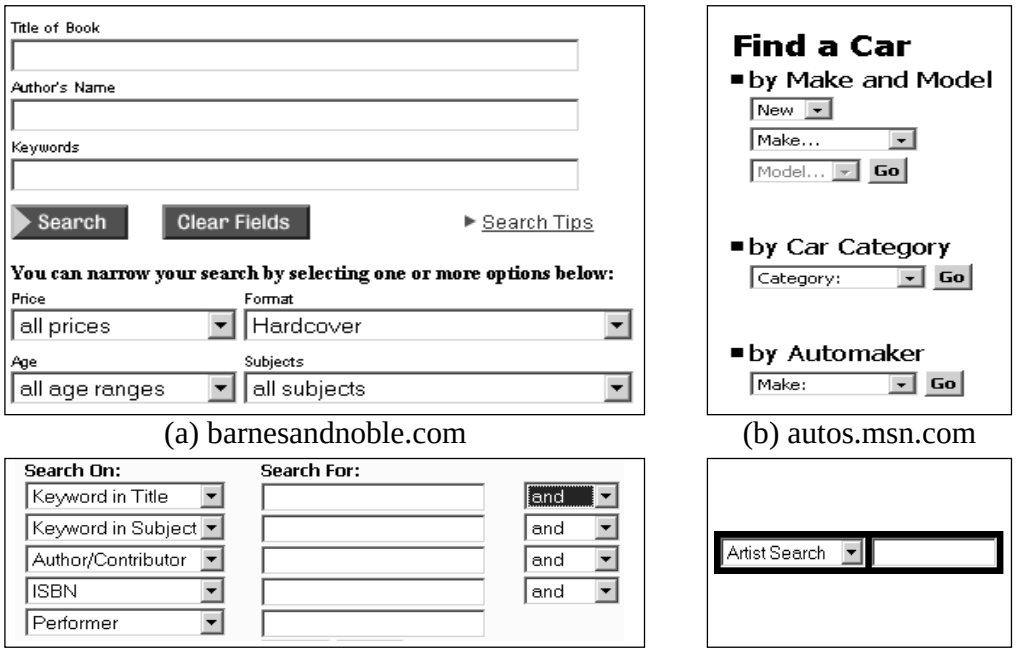
We found that the concerted-complexity behavior seems pervasive on the deep Web. Our survey found that this vocabulary again reveals some concerted structures. There are only 25 constraint patterns overall– which is surprisingly small as a vocabulary for online queries. Figure 15(c) shows several frequently-used patterns: *e.g.*, Pattern 1 is often used to search for keywords contained in a textual attribute (*e.g.*, [author *contains* `"knuth"`]), and Patter 2 expresses a condition for selecting among multiple enumerated values (*e.g.*, {`"round trip"`, `"one way"`}). The distribution is again extremely non-uniform: Figure 15(b) ranks the patterns according to their frequencies (and omits 4 rare attributes in the tail, which occur only once), for each domain and overall. We observe again a characteristic Zipf-distribution, which confirms that a small set of top-ranked patterns will dominate.

We also observe the convergence of constraint vocabularies, both within and across domains. Figure 15(a) summarizes the occurrences of patterns. (To simplify, it similarly omits the rare "only-once" patterns.): The figure marks (x, y) with a "+" if pattern y occurs in source x. Like Figure 12(a), as more sources are seen (along the x-axis), the growth (along y) of the vocabulary slows down and thus the curve flattens rapidly.

However, the constraint vocabulary is more "universal" than the schema counterpart. Unlike Figure 12(a), we observe that the convergence generally spans across different domains (*i.e.*, Automobiles and Airfares are mostly reusing the patterns from Books), which indicates that most constraint patterns are quite generic and not domain specific. Put in a different way, constraint patterns form no localities– we do not observe "dense triangles" in 15(a), unlike in Figure 11. The observation might suggest that "semantics" (*e.g.*, schema attributes) is likely domain-specific, while "syntax" (*e.g.*, constraint patterns) may be more uniform across domains.

(Q11) *How complex are possible queries?* As queries are now formulated through query "forms" (unlike arbitrary SQL queries), is querying becoming "trivial" on the deep Web? To address this question, we examined each source manually for its *maximum query expressiveness*. To ensure "maximum," we need to find the most "advanced" query forms in a source. Thus, for each source, we manually searched the pages within 3 hops from the root page for such a query form. (Our experience and preliminary experiments show 3 hops are sufficient to find advanced query interfaces if they exist at all.)

We measured the *expressiveness* of a query in two dimensions: First, we counted the number of constraints allowed in a query–



(a) barnesandnoble.com (b) autos.msn.com

(c) booksinprint.com (d) coolsiteofday.com

Figure 14: Example query interfaces.

the larger, the more complex a query is. Second, we identified the types of connectives between constraints. We distinguished three different types: 1) conjunctive where queries are constructed by conjunction of constraints (*e.g.*, constraints shown in Figure 14(a) *barnesandnoble.com*); 2) disjunctive where constraints can be combined by *both* or and and (*e.g.*, constraints shown in Figure 14(c) *booksinprint.com*); 3) exclusive where there are multiple constraints but not all of them can used together (*e.g.*, Figure 14(b) *autos.msn.com*).

Figure 16(a) reports the average number of constraints allowed in a query for each domain, and Figure 16(b) the frequency distribution of the three connectives. The average number of constraints are larger than 4 across all domains. The result shows that, while sources tend to share a small number of constraint patterns [Q10], they often allow complex large queries to be constructed.

5. IMPLICATIONS

Sections 3 and 4 reported our observations– Then, what are the likely implications? To further interpret the findings, we discuss our conjectures. They are certainly our own– We believe that, while necessarily subjective, these implications are well motivated by the observations, and might help shape our research directions and solutions for exploring and integrating the deep Web.

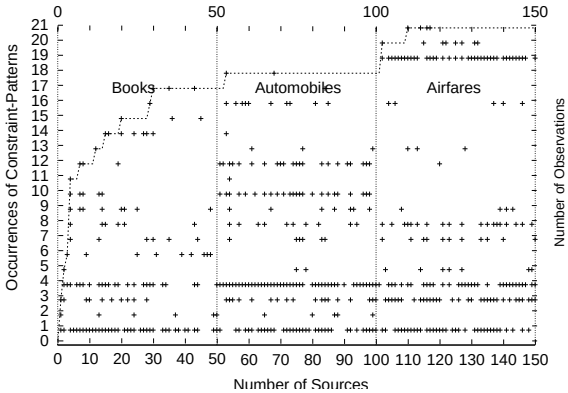
(I1) *Large scale integration is a real and pressing challenge.* While information integration has been actively studied, scalability has not been a main objective. Our community has observed the scalability limitations [7, 8, 9] of current techniques– As sources are proliferating, the deep Web has only made this challenge become real and concrete. Can our techniques "integrate," in a broad sense, heterogeneous Web sources on the order of magnitude of 10^5 [Q2]? With the limited coverage of current directory services, such need seems tantalizing [Q4].

Large-scale integration also implies new problems, such as building a deep Web "search engine" for automatic source discovery, modeling, and selection, beyond more traditional issues of query mediation and data mapping. Some of these issues have been similarly studied for *meta-search* for text databases [10, 11, 12, 13, 14]. The $3.4 : 1$ relative prevalence of structured sources on the Web urges more attentions on these issues. [Q3]

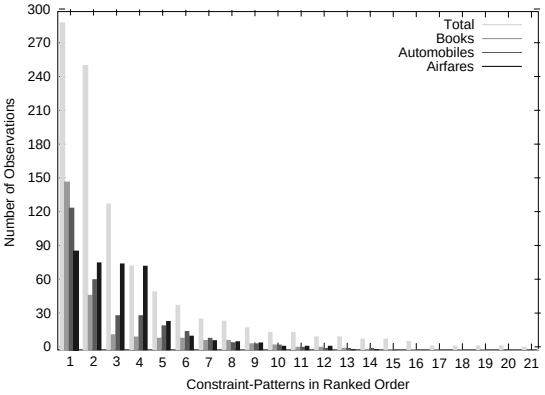
(I2) *Dynamic and ad-hoc integration becomes necessary.* As large scale [I1] also entails, integration will desirably and perhaps even necessarily be *dynamic* and *ad-hoc*: Imagine users of our envisioned deep Web "search engine": Each query will dynamically select various ad-hoc sources (*e.g.*, consider Amy's three queries in Section 1). Such dynamic nature had not been so real before: the research community has mostly focused on traditional scenarios of static systems where sources and mediators are configured *a priori* for specific tasks (say, Books comparison shopping); *e.g.*, [15, 9] survey two main configuration schemes, "global" or "local" as views, for such environments. However, on the deep Web, the new challenges of dynamism and ad-hocness will likely imply hard problems– such as ad-hoc query translation to access new sources without pre-configured semantic annotations. (Is it even possible?)

(I3) *Crawling techniques for source discovery are likely to be different from surface-Web crawlers.* Large-scale integration needs to start with discovering and indexing sources. Specifically, the scale and diversity of Web sources call for automatic 'crawlers" [Q2,5] (while more precise, manual compilation is unlikely to scale, as witnessed by the coverage of directory services [Q4]). Such a crawler will likely be different from that for surface Web– For instance, query interfaces tend to be shallow [Q1] in a site, motivating a site-based shallow crawling (which focuses on promising sites and combs only their top-level pages). The crawler must be more sophisticated to "understand" a query form and extract its key parameters.

(I4) *The deep Web is not entirely hidden– The hiddenness varies across domains.* Link-based navigational access, if available, will surface the deep Web content and thus blur its distinc-



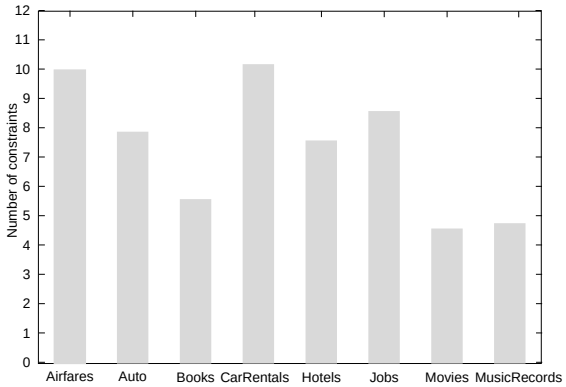
(a) Vocabulary growth over sources.



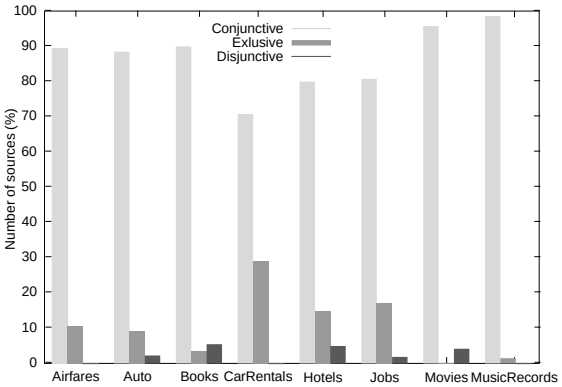
(b) Frequencies over ranks.

(c) Instances of some frequent constraint patterns.

Figure 15: Query vocabulary: Constraint patterns as building blocks for query interfaces.



(a) Average numbers of constraints.



(b) Type of constraint connectives.

Figure 16: Complexity of queries.

tion from the current surface Web. As [Q6] concluded, such hiddenness varies across domains– While some domains (*e.g.*, Airfares and CarRental) remain rather "closed" to be accessed only through query interfaces, others (*e.g.*, Books, Movies) tend to be open, providing browse interfaces as alternative access paths. For such "open" domains, while query-based access will remain important, we can also leverage navigational interfaces (which are more "crawler friendly") in enabling data access and integration.

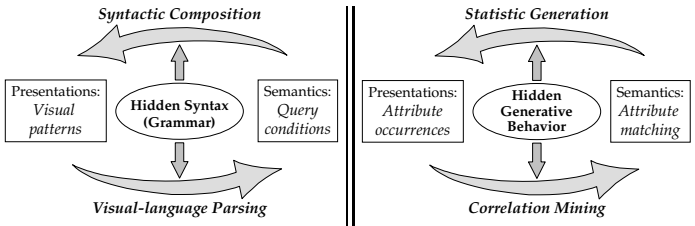
(I5) *Structure-based integration will be essential.* What is the key "semantics" for Web source integration? In meta-search over text sources, *subject topics* are naturally the central notion. In contrast, for structured sources, the notion of *schema* (embedded in queries and results) is clearly essential. As we observed, structured sources dominate on the Web [Q3], their structures are characteristic [Q9], and such structures can often be easily acquired, say, from query interfaces [Q9]. Thus, structured-based integration will likely be both essential and promising for Web databases.

(I6) *Query mediation remains necessary and challenging.* Query mediation has been a traditional focus for integrating heterogeneous sources (*e.g.*, [15]); the problem remains for Web sources: While browse interfaces may sometimes be available, such availability is rather domain-dependent and not universal [Q6]. Query interfaces thus are clearly the primary "access entrance" to sources, universally supported. To avoid online querying, can we take a "warehousing" approach to crawl data offline from sources? While possible, the poor coverage of search-engine caching [Q7] indicates such warehousing unattractive for the deep Web.

Online querying through query forms does not trivialize this "art" [Q11]. The complexity of Web query forms (in terms of number of constraints or attributes) reveals that query mediation in this context does not get much easier. The common focus on conjunctive queries (*e.g.*, [15]) seems well justified by their prevalence.

(I7) *Holistic integration holds promises.*

Our survey apparently indicates *dual* phenomena that together uniquely characterize the deep-Web frontier: First, as a *challenge*: Sources online are virtually unlimited; even for a specific domain of interest, there is often an overwhelming number of alternative sources (the *proliferating sources* phenomenon) [Q2]. Thus, large-scale integration is a real challenge [I1]. Second, as an *opportunity*: However, while sources proliferate, in aggregate, their complexity tends to be "concerted," revealing some underlying "structure." In particular, we observe such concerted structure on the attribute vocabularies [Q9] and query patterns [Q10] across Web sources. Such aggregate vocabularies are clustering in localities and converging in sizes.



(a) query-interface extraction **(b)** query-schema matching

Figure 17: Holistic integration: Exploring regularity.

The dual phenomena seems to hint at a "holistic" approach for integration. By *holistic*, we mean to pursue integration at a large scale and take a holistic view to account for *many* sources together in integration, by globally exploiting observable clues across all sources for resolving the underlying "semantics" of interest– The concerted structure, or the hidden regularity, will likely provide such global clues for semantics discovery. Thus, holistic integration is to apply certain "reverse analysis" for discovering semantics from the observable clues.

For instance, as initial "evidences" for such holistic integration, we have applied this insight to two integration tasks: First, for *query-interface extraction*, as [16] reported, the observation of concerted query patterns motivates us to hypothesize the existence of *hidden syntax*– Such hidden syntax explains the regularity observed. Specifically, we conceptually hypothesize that, as Figure 17(a) shows, the hypothetical syntax (as *hidden regularity*) guides a syntactic composition process (as *connection*) from query conditions (as *semantics*) to their visual patterns (as *presentations*). This hidden syntax effectively transforms the problem: We view query interfaces as a *visual language*; their extraction is precisely the reverse analysis– or *visual-language parsing*.

Second, for *query-schema matching*, as [17, 18] reported, the observation of converging attributes leads us to hypothesize a hidden generative behavior, which probabilistically generates, from a finite vocabulary, the schemas we observed– Such generative behavior explains the regularity observed. As Figure 17(b) shows, the hidden generative behavior (as *hidden regularity*) guides a statistic generation process (as *connection*) from attribute matching (as *semantics*) to their occurrences in interfaces (as *presentations*). This generative behavior constrains how attributes may occur in interfaces–*e.g.*, grouping attributes tend to positively co-occur while synonym attributes negatively. The *reverse analysis* to find attribute matchings is thus the "mining" of correlated attributes, and thus a *correlation mining* approach.

We believe such holistic integration promising for large scale integration– by essentially leveraging the challenge of scale as an opportunity, with two main advantages. First, scalability: By integrating a large number of sources holistically, rather than individually or pairwise, we will be able to cope with the scale of integration. Second, solvability: The large scale can itself be a crucial leverage to solve integration tasks. The holistic approach can take advantage of the large scale (with sufficient "samples") for identifying hidden regularities and applying principled holistic analysis.

6. RELATED WORK

As this paper surveys structured databases on the Web, we discuss several closely related areas. (Note that Section 1 discussed related Web characterization surveys.) Traditionally, information integration (for structured, relational sources) has mainly focused on relatively small-scaled pre-configured systems [15, 9] (*e.g.*, Information Manifold [19], TSIMMIS [20], Clio [21]). Since our interest is the large scale integration of databases on the Web, we will focus on works related to this area. In particular, we discuss text and structured databases integration, for large scale scenarios.

First, for text databases, there has been much effort in large scale distributed "meta-search" (*e.g.*, [22]). Research in this area focuses on constructing "models" for source characterization (*e.g.*, [11]), database selection for query routing (*e.g.*, [14]), collection fusion for merging ranks from different databases (*e.g.*, [23]).

Second, although structured databases dominate on the Web as we surveyed, relatively less work has been done for large scale integration of such sources, as compared with text databases. The same challenges (as we discussed above for text databases), which are equally important and difficult (if not more), exist for structured databases. Some techniques have been proposed to address such challenges: Reference [24] proposes techniques for modeling the query capability of interactive Web sources. Reference [25] introduces an approach for crawling Web databases. References [26, 27, 28] discuss data extraction techniques (or "wrappers"), targeting at HTML pages generated by backend databases.

7. CONCLUSION

This paper presents our survey of databases on the Web, or the so called "deep Web." Our survey was motivated by issues related to exploring and integrating these massive networked databases. On one hand, our "macro" study surveys the deep Web at large, adopting the random IP-sampling approach, with one million samples. We found that the deep Web measured 450,000 Web databases, among which 348,000 were structured. The current representative directory service covered a mere 15.6% of these databases. On the other hand, our "micro" study surveys source-specific characteristics over 441 sources in eight representative domains. We found that deep-Web sources were not entirely hidden– Such hiddenness was domain dependent. Overall, the representative search engine covered only 5% fresh data from these sources. We also observed several interesting "concerted complexities" across deep-Web sources.

We conclude with several implications which, while necessarily subjective, might help shape research directions and solutions. Our main conclusions are– 1) in terms of *problems*: large-scale integration is a real challenge, which likely will mandate dynamic and ad-hoc integration requirements; and 2) in terms of *solutions*: holistic-integration approaches, which discover integration semantics by globally exploiting shallow clues across many sources, are likely to be a key technique for enabling large-scale integration.

8. REFERENCES

[1] BrightPlanet.com. The deep web: Surfacing hidden value. Accessible at `http://brightplanet.com`, July 2000.

[2] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[3] Ed O'Neill, Brian Lavoie, and Rick Bennett. Web characterization. Accessible at `"http://wcp.oclc.org"`.

[4] GNU. wget. Accessible at `"http://-www.gnu.org/software/wget/wget.html"`.

[5] Kevin Chen-Chuan Chang, Bin He, Chengkai Li, and Zhen Zhang. The UIUC web integration repository. Computer Science Department, University of Illinois at Urbana-Champaign. http://metaquerier.cs.uiuc.edu/repository, 2003.

[6] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts, 1949.

[7] William W. Cohen. Some practical observations on integration of web information. In *WebDB (Informal Proceedings)*, pages 55–60, 1999.

[8] Marti A. Hearst. Trends & controversies: Information integration. *IEEE Intelligent System*, 13(5):12–24, September 1998.

[9] Daniela Florescu, Alon Y. Levy, and Alberto O. Mendelzon. Database techniques for the world-wide web: A survey. *SIGMOD Record*, 27(3):59–74, 1998.

[10] Panagiotis G. Ipeirotis, Luis Gravano, and Mehran Sahami. Probe, count, and classify: Categorizing hidden web databases. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, Santa Barbara, Ca., May 2001.

[11] James P. Callan, Margaret Connell, and Aiqun Du. Automatic discovery of language models for text databases. In *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 479–490, Philadelphia, Pennsylvania, USA, June 1999. ACM Press.

[12] David Hawking and Paul B. Thistlewaite. Methods for information server selection. *ACM Transactions on Information Systems*, 17(1):40–76, 1999.

[13] Atsushi Sugiura and Oren Etzioni. Query routing for web search engines: architecture and experiments. In *Proceedings of WWW9*, 2000.

[14] Weiyi Meng, King-Lup Liu, Clement T. Yu, Xiaodong Wang, Yuhsi Chang, and Naphtali Rishe. Determining text databases to search in the internet. In *Proceedings of 24th International Conference on Very Large Data Bases*, pages 14–25, New York City, New York, USA, August 1998. Morgan Kaufmann.

[15] Jeffrey D. Ullman. Information integration using logical views. In *Proceedings of the 6th International Conference on Database Theory*, Delphi, Greece, January 1997. Springer, Berlin.

[16] Zhen Zhang, Bin He, and Kevin Chen-Chuan Chang. Understanding web query interfaces: Best effort parsing with hidden syntax. In *SIGMOD Conference*, 2004.

[17] Bin He and Kevin Chen-Chuan Chang. Statistical schema matching across web query interfaces. In *SIGMOD Conference*, 2003.

[18] Bin He, Kevin Chen-Chuan Chang, and Jiawei Han. Discovering complex matchings across web query interfaces: A correlation mining approach. In *SIGKDD Conference*, 2004.

[19] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proceedings of the 22nd VLDB Conference*, pages 251–262, Bombay, India, 1996. VLDB Endowment, Saratoga, Calif.

[20] Yannis Papakonstantinou, Héctor García-Molina, and Jeffrey Ullman. Medmaker: A mediation system based on declarative specifications. In *Proceedings of the 12th International Conference on Data Engineering*, New Orleans, La., 1996.

[21] Renée J. Miller, Mauricio A. Hernández, Laura M. Haas, Lingling Yan, C. T. Howard Ho, Ronald Fagin, and Lucian Popa. The Clio project: managing heterogeneity. *SIGMOD Rec.*, 30(1):78–83, 2001.

[22] Luis Gravano, Chen-Chuan K. Chang, Héctor García-Molina, and Andreas Paepcke. STARTS: Stanford protocol proposal for internet retrieval and search. Accessible at `http://www-db.stanford.edu/~gravano/starts.html`, August 1996.

[23] Luis Gravano and Héctor García-Molina. Merging ranks from heterogeneous internet sources. In *Proceedings of 23rd International Conference on Very Large Data Bases*, pages 196–205, Athens, Greece, August 1997. VLDB Endowment, Saratoga, Calif.

[24] Bertram Ludäscher and Amarnath Gupta. Modeling interactive web sources for information mediation. In *Advances in Conceptual Modeling: ER '99 Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling, Paris, France, November 15-18, 1999, Proceedings*, volume 1727 of *Lecture Notes in Computer Science*, pages 225–238. Springer, 1999.

[25] Sriram Raghavan and Hector Garcia-Molina. Crawling the hidden web. In *Proceedings of 27th International Conference on Very Large Data Bases*, Roma, Italy, 2001. Morgan Kaufmann.

[26] James Caverlee, Ling Liu, and David Buttler. Probe, cluster, and discover: Focused extraction of qa-pagelets from the deep web. In *ICDE Conference*, 2004.

[27] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In *The VLDB Journal 2001*, pages 109–118, 2001.

[28] D. W. Embley, Y. Jiang, and Y.-K. Ng. Record-boundary discovery in Web documents. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 467–478, Philadelphia, Pennsylvania, USA, June 1999.