CHAPTER 1

Search Based Applications

1.1 INTRODUCTION

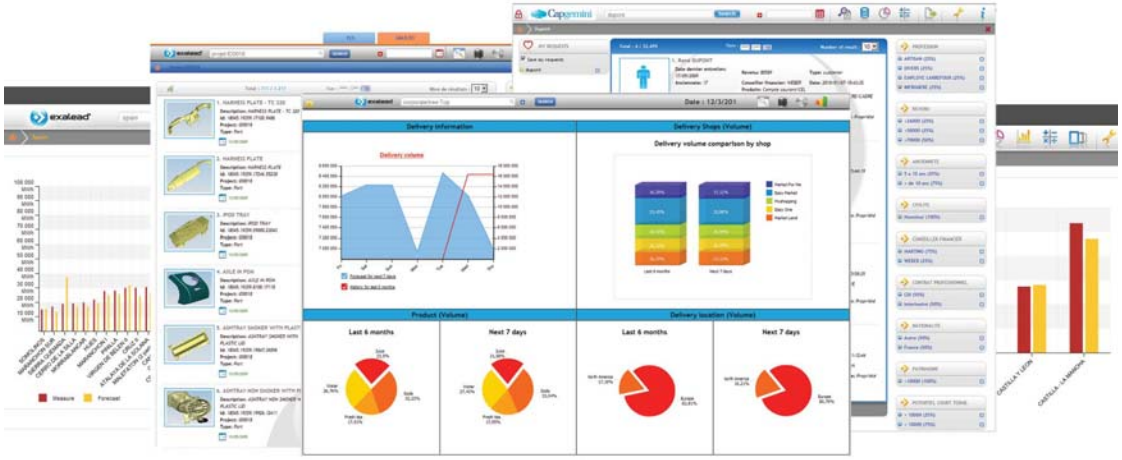


Figure 1.1: Can you see the search engine behind these screens?

Management of information via computers is undergoing a revolutionary change as the frontier between databases and search engines is disappearing. Against this backdrop of nascent convergence, a new class of software has emerged that combines the advantages of each technology, right now, in Search Based Applications.

Until just a short while ago, the lines were still relatively clear. Database software concentrated on creating, storing, maintaining and accessing structured data, where discrete units of information (e.g. *product number*, *quantity available*, *quantity sold*, *date*) and their relation to each other were well defined. Search engines were primarily concerned with locating a document or a bit of information within collections of unstructured textual data: short abstracts, long reports, newspaper articles, email, Web pages, etc. (classic Information Retrieval, or IR; see Chap. 3).

Business applications were built on top of databases, which defined the universe of information available to the end user, and search engines were used for IR on the Web and in the enterprise.

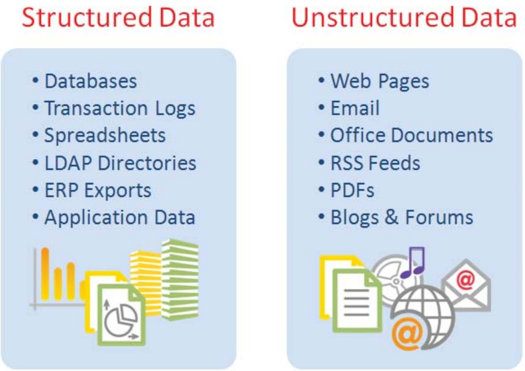


Figure 1.2: Databases have traditionally been concerned with the world of structured data; search engines with that of unstructured data (some of these data types, like HTML pages and email messages, contain a certain level of exploitable structure, and are consequently sometimes referred to as "semi-structured").

Such neat distinctions are now falling away as the core architectures, functionality and roles of search engines and databases have begun to evolve and converge. A new generation of non-relational databases, which shares conceptual models and structures with search engines, has emerged from the world of the Web (see Chapter 4), and a new breed of search engine has arisen which provides native functionality akin to both relational and non-relational databases (described in Chapters 3-9 and listed in Chapter 10).

It is this new generation engine that supports Search Based Applications, which offer precise, multi-axial information access and analysis that is virtually indistinguishable at a surface level from database applications, yet are endowed with the usability and massive scalability of Web search.

1.1.1 WHAT IS A SEARCH BASED APPLICATION?

We define a *Search Based Application* (SBA) as any software application built on a search engine backbone rather than a database infrastructure, and whose purpose is not classic IR, but rather mission-oriented information access, analysis or discovery.[1]

[1] This new type of application has alternately been referred to as a "search application," "search-centric application," "extended business application," "unified information access application" and "search-based application." The latter is the label used by IDC's Susan Feldman, one of the first industry analysts to identify SBAs as a disruptive trend and an influential force in the SBA label being adopted as the industry standard. Feldman has recently moved toward a more precise definition, limiting SBAs to "fully packaged applications" supplying "all the tools that are commonly needed for a specific task or workflow," that is to say, commercial-off-the-shelf (COTS) software [Feldman and Reynolds, 2010]. However, we prefer a broader definition to underscore one of the great benefits of the SBA model: the ability for anyone to rapidly and inexpensively develop highly specific solutions for unique contexts, and, following the same pattern as database applications, we expect both custom and COTS SBAs to flourish over the next decade.

Definition: Search Based Application

A software application that uses a search engine as the primary information access backbone, and whose main purpose is performing a domain-oriented task rather than locating a document. Examples:

Customer service and support
Logistical track and trace
Contextual advertising
Decision intelligence
e-Discovery

SBAs may be used to provide more intuitive, meaningful and scalable access to the content in a single database, hiding away the complexity of the database structure as data is extracted and re-purposed by search engine techniques. They may also be used to autonomously and intelligently gather together massive volumes of unstructured and structured data from an unlimited number of sources (internal or external) and to make this aggregate data available in real time to a wide base of users for a broad range of purposes.

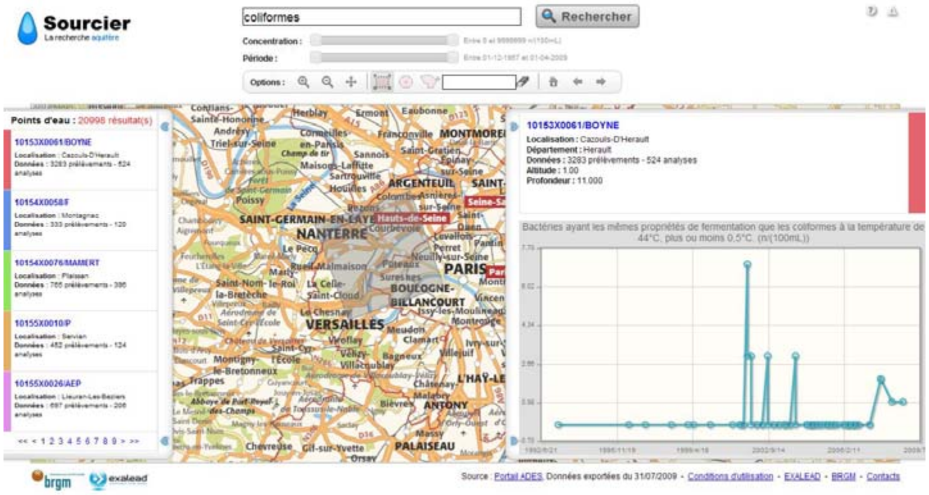
While search engines in the SBA context complement rather than replace databases, which remain ideal tools for many types of transaction processing, this 're-purposing' of search engines nonetheless represents a major rupture with a 30-year tradition of database-centered software application development. In spite of the significance of this shift, the SBA trend has been unfolding largely under the radar of researchers, systems architects and software developers. However, SBAs have begun to capture the focused attention of business.[2]

> "The elements that make search powerful are not necessarily the search box, but the ability to bring together multiple types of information quickly and understandably, in real time, and at massive scale. Databases have been the underpinning for most of the current generation of enterprise applications; search technologies may well be the software backbone of the future."
>
> —Susan Feldman, IDC LINK, June 9, 2010

[2] SBAs are fueling a significant portion of the growth in the search and information access market, which IDC estimates grew at double digit rates in 2007 and 2008, and at a healthy 3.9% (to $2.1 billion) in 2009 [Feldman and Reynolds, 2010]. Gartner, Inc. estimates an compound annual growth rate of 11.7% from 2007 to 2013 for the enterprise search market [Andrews , 2010].

1.2 HIGH IMPACT, LOW RISK SOLUTION FOR BUSINESSES

SBAs offer businesses a rapid, low risk way to eliminate some of the peskiest and most common information systems (IS) problems: siloed data, poor application usability, shifting user requirements, systemic rigidity and limited scalability.



coliformes

Points d'eau : 20998 résultat(s)

10153X0061/BOYNE
Localisation : Cazouls-D'Herault
Données : 3283 prélèvements - 524 analyses

10154X0058/F
Localisation : Montagnac
Données : 333 prélèvements - 120 analyses

10154X0076/MAMERT
Localisation : Plaissan
Données : 765 prélèvements - 386 analyses

10155X0010/P
Localisation : Servian
Données : 452 prélèvements - 124 analyses

10155X0026/AEP
Localisation : Lieuran-Les-Beziers
Données : 697 prélèvements - 206 analyses

<< < 1 2 3 4 5 6 7 8 9 > >>

10153X0061/BOYNE
Localisation : Cazouls-D'Herault
Département : Herault
Données : 3283 prélèvements - 524 analyses
Altitude : 1.00
Profondeur : 11.000

Bactéries ayant les mêmes propriétés de fermentation que les coliformes à la température de 44°C, plus ou moins 0,5°C. (n/(100mL))

Source : Portail ADES Données exportées du 31/07/2009 - Conditions d'utilisation - EXALEAD - BRGM - Contacts

Figure 1.3: Search engine-based *Sourcier* makes vast volumes of structured water quality data accessible via map-based search and visualization, and ad hoc, point-and click-analysis.

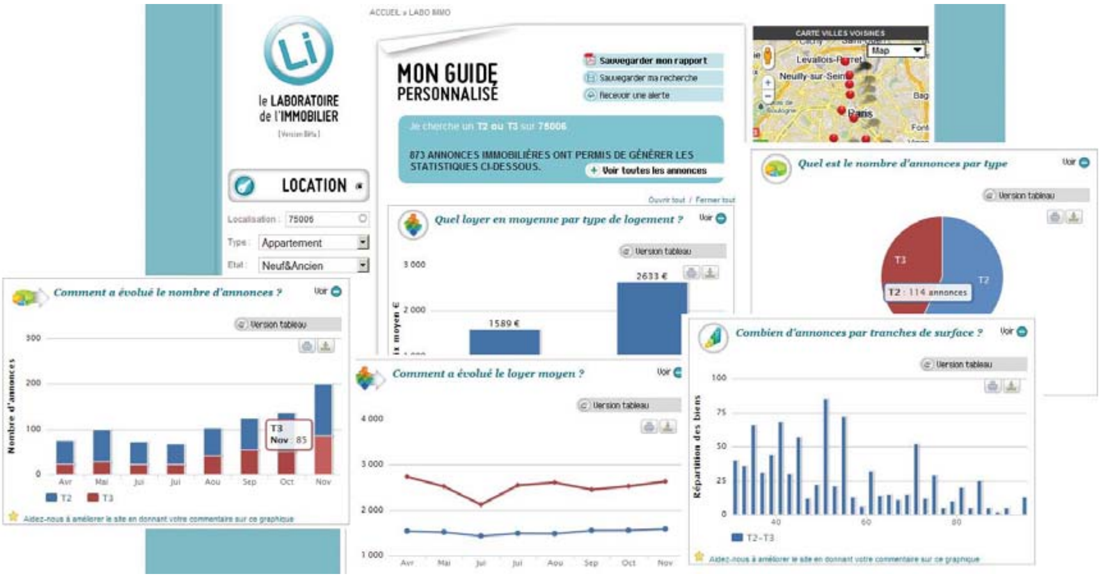
Even though SBAs allow business to clear these hurdles and bring together large volumes of real time information in an immediately actionable form—thereby improving productivity, decision making and innovation—too many in the business community are still unaware that search engines can serve as an information integration, discovery and analysis platform. This is the reason we have written this book.

1.3 FERTILE GROUND FOR INTERDISCIPLINARY RESEARCH

We have also undertaken this project to introduce SBAs to a wider segment of the data management research community. Though the convergence of search and database technologies is gradually being recognized by this community[3], many researchers are still unaware of the pragmatic benefits of SBAs and the mutually beneficial evolutions underway in both search and database disciplines.

[3] See, for example, recent workshops like *Using Search Engine Technology for Information Management* (USETIM'09) that was held in August 2009 at the *35th International Conference on Very Large Data Bases* (VLDB09), which examines whether search engine technology can be used to perform tasks usually undertaken by databases. `http://vldb2009.org/?q=node/30`

However, as a group of prominent database and search scientists recently noted, exploding data volumes and usage scenarios along with major shifts in computing hardware and platforms have resulted in an "urgent, widespread need for new data management technologies," innovations that will only come about through interdisciplinary research.[4]



Je cherche un T2 ou T3 sur 75006

873 ANNONCES IMMOBILIÈRES ONT PERMIS DE GÉNÉRER LES STATISTIQUES CI-DESSOUS.

Localisation : 75006

Type : Appartement

Etat : Neuf&Ancien

Figure 1.4: This Akerys portal generates personalized, real-time real estate market intelligence based on unstructured online classifieds and in-house databases.

1.4 A VALUABLE TOOL FOR DATABASE ADMINISTRATORS

Like their research counterparts, many Database Administrators (DBAs) are also unfamiliar with SBAs. We hope this book will raise awareness of SBAs among DBAs as well, because SBAs offer these professionals a fast and non-intrusive way to offload overtaxed systems[5] and to reveal the full richness of the data those systems contain, opening database content up for free-wheeling discovery and analysis, and enabling it to be contextualized with external Web, database and enterprise content.

[4] From the *The Claremont Report on Database Research*, the summary report of the May, 2008 meeting of a group of leading database and data management researchers who meet every five years to discuss the state of the research field and its impacts on practice: `http://db.cs.berkeley.edu/claremont/claremontreport08.pdf`

[5] *Offloading a database* means extracting all the data that a user might want to access and indexing a copy of this information in a search engine. The term *offloading* refers to the fact that search requests no longer access the original database, whose processing load is hence reduced.

1.5 NEW OPPORTUNITIES FOR SEARCH SPECIALISTS

For search specialists who are not yet familiar with SBAs, we hope to introduce them to this significant new way of using search technology to improve our day-to-day personal and professional lives, and to make them aware of the new opportunities for scientific advancement and entrepreneurship awaiting as we seek ways to improve the performance of search engines in the context of SBA usage.

1.6 NEW FLEXIBILITY FOR SOFTWARE DEVELOPERS

We also hope to make software developers aware of the new options SBAs offer: one doesn't always need to access an existing database (or create a new one) to develop business applications or to meticulously identify all user needs in advance of programming, and one need not settle for applications that must be modified every time these needs or source data change.

1.6.1 LECTURE ROADMAP

While this diversity of audiences and the short format of the book necessitate a surface treatment of many issues, we will consider our mission accomplished if each of our readers walks away with a solid (if basic) understanding of the significance, function, capabilities and limitations of SBAs, and a desire to go forth and learn more.

To begin, we'll first take a look at the ways in which information access needs have changed, then provide a comparative view of ways in which search engines and databases work and how each has evolved. We'll then explain how SBAs work and how and when they are being used, including presenting several case studies.

Finally, we will situate this shift within the larger context of evolutions taking place on the Web, including conceptions of the Deep Web, the Semantic Web, and the Mobile Web, and what these evolutions may mean for the next generation of SBAs.