Oracle Text

An Oracle White Paper
March 2002
EXECUTIVE OVERVIEW

Oracle Text adds powerful text search and intelligent text management capabilities to Oracle9i. Oracle Text indexes any document or other textual content to deliver fast, accurate retrieval of information. Oracle Text is an invaluable infrastructure component for a host of applications – web-based content management, e-business catalogs, portals, and other online text information sources in multiple languages. Oracle Text has proven search quality and the best globalization features in the market. Users of the Oracle9i database need look no further than Oracle Text for intelligent management of all their text information.

INTRODUCTION

Over the last decade, organizations have invested heavily in reliable and rapid access to structured (business) data stored in database systems. However, this data represents a fraction of all corporate information. A far larger volume of data exists as text in documents, web pages, manuals, reports, email, faxes, and presentations. These valuable sources of strategic business information are often inaccessible, and often not managed in an effective manner. Users accessing organizations’ information – whether they are employees visiting an intranet portal or buyers browsing a catalog – need support from an underlying text infrastructure to find what they want.

The reality today is that organizations can warehouse, analytically process, slice - and - dice the 20% of corporate data that comprises of numbers and dates, but the rest that sits as text – in emails, files, annotations to database records - often goes untapped. Important text assets stay hidden, and the latent information they carry remains obscure to decision makers. Knowledge workers looking for information find search quality over such assets to be poor. The poor quality of decisions, as well as the loss of productivity contributes to higher expenses. As companies embrace the web and create globally interconnected systems drawing information from many (often unstructured) sources, the problem is aggravated.
Unlocking the value of an organization’s textual information has been a long-term challenge. Historically, text has been perceived to require a different set of technologies for retrieval and management than structured data. This misperception has burdened organizations with multiple storage systems and development environments, and has stood in the way of effectively integrating all corporate information assets.

As a legacy of this misperception, many companies today buy different products for solving their text searching needs and their structured data (database) searching needs. Not only is this approach costly over the life cycle of purchasing, integrating, operating and maintaining different products, but it also results in poor performance, lack of an integrated view of information assets, and a high latency in development of applications. Further, purveyors of specialty servers can seldom deliver the high reliability, throughput and multi-platform scalability that an enterprise database platform delivers.

What if it were possible to extend the power and advantages of enterprise database systems to all corporate information, including text and other unstructured data? After all, text data is real data that warrants the infrastructure of a real database and proven tools for application development. In this white paper, we look at such an approach in the form of Oracle Text.

ORACLE TEXT

Oracle Text (formerly known as interMedia text) offers a complete text search solution. Oracle Text is included with both the Oracle9i Standard and Enterprise Editions. For users of an Oracle database, Oracle Text eliminates the need to evaluate, buy and integrate a separate text searching product.

Oracle Text provides specialized text indexes for traditional full text retrieval applications – such as website searching, e-business catalogs applications, document classification and routing applications, text warehousing, document libraries and archives, and so on.

Oracle Text provides the best text search quality in the industry. This proven Oracle technology includes patented features for intelligently processing search terms to return the right information – apparent or latent – from the underlying data sources. Oracle Text excels at performing exact and inexact matches, word positioning comparisons, intelligent match, high-accuracy relevance ranking of returned results, XML searches, and so on.

Oracle Text can filter and extract content from all the different document formats commonly used in organizations. It supports more than 150 document formats including popular ones like the Microsoft Office™ file formats, the Adobe PDF™ family of formats, as well as HTML and XML.
Oracle Text offers the best set of multilingual features in the market - supporting search across documents in western languages (English, French, Spanish, German, etc.) as well as Japanese, Korean, Traditional and Simplified Chinese.

As part of the Oracle9i database, Oracle Text transparently integrates with and benefits from a number of key enterprise features such as

- Data partitioning (for higher throughput and availability)
- Real application clustering (for the highest server scalability)
- Query optimization (to ensure the best response time, not only for pure text queries, but also ‘mixed’ queries that combine text search with structure database search)
- Tools and development environments (to increase developers’ productivity)
- Administration and manageability (to reduce operational costs and ensure ease of maintenance)
- Integrated security (to protect all your information assets with the same rigor as your database data)

These aspects of integration are also greatly beneficial to system and database administrators, who do not have to undergo a paradigm shift to learn to manage and organization’s text assets. Oracle Text brings together the worlds of the librarian and the DBA into a single integrated product.

**USING ORACLE TEXT**

Oracle Text search technology provides the foundation for text processing in the Oracle Internet File System, Oracle Ultra Search, the Oracle eBusiness Suite, and Oracle9i Application Server Portal (Oracle9iAS Portal), as well as the search technology on Oracle.com. Oracle uses its own Text technology across the board.

Generally, Oracle Text is targeted for three broad segments.

**Web Site Searching**

Corporate websites are offering access to a vast amount of information – products, news, jobs, partners, etc. (Figure 1) to both internal and external users. Organizations are also transforming traditional file based websites into database-backed ones. Such database-backed websites provide a number of advantages:

- They can be administered – secured, backed up, restored – and queried just like other enterprise resources.
They are transactional – the database ensures that any modifications that fail halfway are ‘rolled back’. Traditional file systems are susceptible to failures and corruption, and it is difficult to maintain consistency and integrity across large distributed sites.

They can be used to dynamically generate and format data.

They scale better over large numbers of users.

With technologies like Oracle Text, all the search capabilities you need are available in your database.

Oracle Text makes it easy to develop and integrate a site search with the other features of a database-backed web site. Sites get a powerful, high quality search in any language.

In recent years, eBusiness companies have moved their Web content into databases. At the same time, legacy data in existing stores and repositories still has to be addressed. For this purpose, Oracle9i introduces Oracle Ultra Search, a web-based application built on top of Oracle Text. Oracle Ultra Search crawls diverse data repositories, such as multiple databases, e-mail/IMAP servers, web servers etc., gathering and storing index information in an Oracle Text index. More than one source can be included in a single index. Additional information on this topic can be found in the Oracle Ultra Search white paper.

Figure 1. Web Site Searching with Oracle Text.
Along with base technology, Oracle also has a higher level solution for Intranet searching. Oracle Text by itself is about content in the database, Ultra Search expands that with a crawling and indexing approach for multiple data stores, using the underlying Text technology. Ultra Search is available packaged with the database as well as with Portal in Oracle iAS. Figure 2 shows a screenshot of Ultra Search.

![Ultra Search Screenshot](Image)

**Figure 2. Oracle Ultra Search**

**eBusiness Catalogs**

eBusiness catalogs are an important part of organizations that reach consumers over the Web. An electronic cataloging record contains information about an item – usually including a textual description of the item, identification, information like a stock or SKU code, as well as other attributes like price, quantity, etc.

If customers can’t find an item, they will go to a competitor’s website. There are different issues in finding the right item in a catalog. In the case of a music CD, a customer may know the artist’s name, but not the latest album name; or perhaps some words of a song but not the exact title. An efficient eBusiness catalog search must address several different ways of finding the right asset (Figure 3).

Text searching is not the only problem. All catalogs contain structured data (numbers, dates) and unstructured data (text). Therefore the ability to retrieve structured and unstructured data in a single query is very important to ensure the
consistency and availability of the items on the website. Oracle Text catalog index was designed for this type of scenario.

Oracle Text introduces the catalog index type that is well-suited for indexing small text fragments and related information in typical eBusiness environments such as an auction website.

Online catalogs are increasingly prevalent, and serve as the vehicle for a number of applications ranging from online retailing, internal purchasing, benefits administration and so on. As items on such catalogs proliferate, it is important to specialize technologies dealing with catalogs so that users can efficiently find what they want.

The Oracle9i catalog index is designed for optimal query performance in situations where textual data (such as an item name) has to be accessed along with non-text data (such as price), and ordered by attributes such as date. For such queries, Oracle’s catalog indexes can deliver response times, which are orders of magnitude better than standard text indexes. Catalog vendors are therefore able to not only service users quickly, but also use less per capita hardware to do so.

Oracle Text scales and performs with the database. It is part of the database and therefore is highly available for eBusiness catalogs.

Figure 3. A Typical Catalog at Oracle Purchasing.
Document Archives and Digital Libraries

Knowledge workers in a variety of industries must access and process large amounts of information to make day-to-day decisions. An oil field engineer in a remote site needs to access all technical information from his truck to solve a problem. A physician needs to access the latest issue of a medical journal from a hospital room or his private office. A legal researcher needs to access to case law, or an editor to past news stories. All these uses require efficient text searches on large amounts of text from diverse sources.

As the amount of information increases, classification and personalization emerge as essential technologies that help users find what they are looking for in an efficient way.

Oracle Text provides classification, personalization, and text mining along with other linguistic technologies. Personalization in a search context is the ability to retrieve text assets given a particular user profile. A salesman searching for “automobile” expects to see different results than an engineer. The personalization capabilities of Oracle Text improve search effectiveness and quality by tying the retrieved results to user-profiles.

Text mining, the text counterpart of traditional data mining, is the process of textual data analysis that leads to the discovery of unknown information. Themes, gists and other features extracted from documents by Oracle Text can be used to mine for latent information.

Case Study: DER SPIEGEL’s DIGAS application

As Germany's leading news magazine and the largest news weekly in Europe, DER SPIEGEL is a market and opinion leader. The SPIEGEL Group generates annual revenues of 691 million marks (around $310 million).

To stay at the top of the news business, DER SPIEGEL reporters and researchers depend on fast access to up-to-date information. This information comes from many sources, in many languages, and must be kept current. To stay informed, DER SPIEGEL maintains a huge digital archive.
DIGAS (DIGital Archiving System) is the application used for research and fact checking in the Research Department and by the journalists writing for DER SPIEGEL, SPIEGEL TV, and SPIEGEL Online. The archive contains more than 12 million documents and about 500,000 pictures. Oracle Text provides the infrastructure for DIGAS’s powerful searching features. Figure 4 shows DIGAS in action.

Figure 4. DIGAS search feature

EVALUATING THE ORACLE TEXT TECHNOLOGY

A number of different approaches to text retrieval and management are available in the market. Oracle Text’s strong points are integration with searches over all other types of data, high performance, as well as superior quality of search.

We can classify all the products in the area of search and text management according to two important metrics: the quality of the search and the level of integration of all types of data (structured and text).

1 Used by permission from DER SPIEGEL.
For anyone who has entered a search keyword only to be rewarded with thousands of obscure hits, the importance of search quality needs no emphasis. Search quality can be quite objectively measured in a number of ways – such as the number of relevant items returned by a search, the ratio between semantically relevant and irrelevant hits, the ordering of returned results and so on.

The *level of integration* metric is an important one since it is tied closely with cost. ‘Specialty’ products which poorly integrate text data with other forms of enterprise data lead to high life-cycle costs – not only in terms of the application-level integration that must be done by the user, but also in terms of low performance and throughput.

Figure 5 shows how Oracle Text compares with other text retrieval technologies for these metrics. Some custom solutions have good search quality but their data management infrastructure is very primitive. Other custom solutions can manage any type of data but the search quality is poor. Web search engines can manage only simple-language keyword searches, text and their average search quality is poor (they return thousands of “hits”). Document management and text only products tend to have good search quality but not very good data integration since their concentration is on text only assets. Other databases and platform vendors provide less than average search quality with average data management. Oracle Text provides high integration of text and structured data with high search quality.

![Figure 5. Comparing the Oracle Text Solution.](image-url)
PROBLEMS WITH MANAGING TEXT AND DATA SEPARATELY

For a number of years, text retrieval was a niche market, and most mainstream organizations were more preoccupied with managing their numbers and dates rather than their documents. A number of exotic technologies grew into this niche market, focusing on the differences between text data and other corporate data, rather than on the similarities.

More recently, the Internet has changed the nature of data management. Companies are moving information and applications to extranet and intranet web sites. The goal is to make information as uniformly accessible as possible. The Internet has led us to expect ad-hoc access to all information, from wherever we are, with a very short response time and at low cost. As the data sources multiply, organizations begin to realize that text data is also a valuable corporate asset and look for regimes to manage it with the same degree of effectiveness and rigor as other information assets such as sales and inventory databases, personnel records, or transaction information.

Over the years, Oracle has established itself as the premier solution for structured data (normal business data). However, organizations have been slow to apply standard database capabilities to the textual data scattered across PCs, file servers, etc. As these organizations face the challenge of utilizing their text information assets, many encounter the specialty text vendors repositioning themselves as knowledge management or portal-search platforms. This launches a debate on whether to install a “special” text search-processing engine.

The attempt to separate text and normal business (structured) data fails because of a number of reasons:

- **Cost**
  
The cost of buying two different products implies upgrades, training, and support fees for each product.

- **Complexity**
  
  Two sets of repositories results in duplication of configuration, administration, and training – increasing overall system complexity (Figure 6).

- **High latency of development and deployment**
  
  There are two different APIs that the development team has to learn. Development of new applications and integration with existing ones takes more time and resources.

- **Performance**
  
  There is inefficient processing of mixed queries (structured business data and text), at great detriment to response time and...
throughput as the results of each query must be combined at the application layer.

**INTEGRATING TEXT AND DATA IN ORACLE**

Oracle Text fully leverages the Oracle9i database. The text engine runs inside the database server. Indexes that enable fast text search reside in the database and are managed in the same way as those on traditional business data. The API to create text applications is the popular SQL language known to developers all over the world. All these deliver high reliability, scalability and platform availability, user satisfaction, and performance – at no extra cost.

In a fully integrated system like Oracle, the text and all its indexes are stored within the relational database (Figure 7). A fully integrated system must use the database kernel to process the entire query. In this way, the text index is treated just like any other native index type.

The following advantages of integration are immediately apparent:

- **Low Cost**
  Oracle Text is part of the Oracle9i Enterprise and Standard Editions. There are no separate products to deal with.

- **Low complexity**

---

Figure 6. Two Different APIs for Accessing Text and Relational Data.
Text is just data, like structured data, and it is easy to develop and integrate search application with existing systems.

- **High Performance**
  The database will choose the fastest plan to execute queries that involve text and structured content.

- **High Integrity**
  Since text is stored in the database it inherits all the integrity benefits – for example, any update to the database can be reflected in the text search functionality, which means users can get an integrated, holistic view of all their data.

- **Superior Manageability**
  Oracle Text can be managed from standard enterprise management tools, leveraging commonly available administrators’ skills. Figure 8 shows the administration of Oracle Text from OEM (Oracle Enterprise Manager).

- **Leveraging existing management skills**
  DBAs do not have to learn any new skills to manage text. They can use the same tools to manage text and structured data.

Figure 7. SQL API for Accessing Text and Relational Data.

In addition to all the integration benefits, Oracle Text stands on its own as the premier text search technology in the industry. We have briefly mentioned issues like search quality before. We now look at some of the aspects that make
Oracle Text – independent of the integration with the database platform – a truly superior technology offering.

![Figure 8. Managing Oracle Text from Oracle Enterprise Manager.](image)

**SEARCH QUALITY**

The world of structured searches is the world of exact matches and precise semantics. An employee with ID 25000 is expected to be different from an employee with ID 25001. The world of text search, is, however, one where imprecision is to be tolerated, and, in some cases, even encouraged. Users express searches differently and expect good search results for answers. For example a user may search for “bookshelf” while another user may look for “book-shelf” or “bookshelves” in a furniture catalog. “Refridgerator” and “refrigerator” are another example. In each case, the search system must be able to return the same results though the queries are entered differently. A quality search service can find the right results even when users make spelling or grammatical errors, use synonyms, or otherwise vary the inputs to search.

A poor search service is expensive. Merchandise does not sell if the users can’t find it. Knowledge workers are less productive if they can’t find the knowledge assets. Vital information stays hidden and additional training has to be given to personnel.

A related problem is that of novice users. The Internet has brought many new computer users who are interacting with information systems for the first time.
From the proverbial grandmother on the Web to the novice employee, users may also express searches in ways that often underscore the differences between their mental models with and those of the system designers. How can people find exactly what they want when they don’t know how to express a query?

Under the covers there is a significant amount of work for the search system. How well the search system solves the problem is a characteristic of its quality. There are a number of strategies that Oracle Text offers to deliver extremely good search quality from the information access aspect as well as the intelligent text management.

Oracle Text can intelligently process search queries using several strategies:

- **Keyword searching.** Searching for keywords in a document. Typically the user enters one or more keywords that best describe the query. For example: “Java”.

- **Context queries.** Searching for words in a given context. Here, the user searches for text that contains words near to each other. For example “Java near business” will order results according to how close “Java” is to “business”.

- **Boolean operations.** Combining keywords with Boolean operations. User can express a query connecting Boolean operations to the keywords. For example “Java and business” – the presence of both words anywhere in a text.

- **Linguistics features.** Using fuzzy and other natural language processing techniques. User search for text that is about something. For example, documents “about Java”.

- **Pattern matching.** Retrieval of text that contains a certain property. User search for text that contains words that contain a string. For example, search for text that contains a word that has “oxy” in it.

Oracle Text also offers a number of advanced intelligent text management features that improve search quality. Using an internal knowledge base that contains a generic classification with relationships among its structures it is possible to find out what a piece of text is about, not just the words it contains. This knowledge base can be extended with a standard thesaurus to include terms from a particular industry – medicine and medical terminologies for example. As part of the text management features, Oracle Text provides mechanisms to automate text classification and grouping of similar documents according to a topic. Some of these capabilities are discussed in the Advanced Features section below.
INTERNATIONALIZATION

In today’s world the ability to communicate the same idea in different languages is a key to success. Global operations generate text data in different languages. The Internet is a global market where companies buy and sell assets from or to any place at any time in any language.

Oracle Text offers the best set of multilingual features for internationalization of search applications. With the rich set of lexers that Oracle Text provides, it is possible to search documents in western languages (English, French, Spanish, German, etc.), Japanese, Korean, Traditional and Simplified Chinese. Cross-language search is also feasible using this technology. Figure 9 shows the use of some of those features. Oracle Text is used to manage text content in dozens of languages in more than a hundred countries around the globe.

ADVANCED FEATURES

State of the art search quality does not guarantee success without advanced features for text management and document services.

Thesaurus

Certain business areas like medicine or pharmaceutical have standards for thesauri. A thesaurus consists of a controlled vocabulary with a structure that...
denotes hierarchy and relationships among the words. Thesauri improve search quality: users can search for related terms, synonyms, etc.

**Knowledge base**

Oracle Text offers an internal knowledge base of half-a-million terms as the foundation for its linguistic engine that can analyze and generate the main themes of a piece of text. These types of features are extremely useful for building classification, routing, and clustering applications.

**Classification and Clustering**

A classification application is one that classifies an incoming set of documents based on their content. Figure 10 shows the structure of a classification application. The action performed after the documents matched certain rules could be delivery via email to subscribers or publication to an internal website.

Classification is a powerful tool for customizing information delivery, automatic content organization and reducing information overload.

![Figure 10. Structure of a Classification Application.](image)

Oracle Text offers the index type CTXRULE and the package CTX_CLS for building classification systems.

The CTX_CLS package generates CTXRULE query rules for a set of documents. The user has to supply a training set consisting of categorized documents and each document must belong to one or more categories. The package generates the queries that define the categories and then writes the results to a table. Figure 11 shows an example of a classification system for the medical domain.
Contrary to classification, clustering is the unsupervised classification of patterns into groups. Oracle Text offers the CTX_CLUSTER package for building clusters. This package is currently available from OTN.

**Document Services**

Powerful document services allow an application to display an HTML version of a common binary format (like PDF among others) with search query terms highlighted. This type of service can be very useful for browsing strategies. Other document services are themes, gists, and query feedback.

**Index Partitioning**

In Oracle9i it is possible to create a text index on a local partition. The major benefits of partitioning are:

- **Manageability.** An administrator can decide how to partition the index, which partitions are online/offline, which partitions to backup, and so on.
- **Performance.** There can be a tremendous improvement in performance with index partitioning.
XML Indexing and Searching

As part of the Oracle platform, Oracle Text provides rich XML text search features that can be used in any XML-driven application. XML (extensible Markup Language) is a standard that can be used for interchanging data among applications. Oracle Text can search XML documents, following path expressions to limit the search to specific portions or nodes of these documents.

This ability to query XML efficiently helps organizations embrace loosely-coupled intranet/extranet-based architectures for managing their information assets.

The ability to combine these special services creates an exceptionally powerful platform. For example, in an eBusiness environment, a particular vendor application can send a product description in XML to a catalog classification system where it will be posted in a product category.

XML Type

Oracle9i introduces a new, native datatype called XMLType, which helps you store XML content directly into the database. Oracle Text has been enhanced to support the text indexing of XMLType documents. This feature allows efficient access to XML documents, and makes it possible for organizations to deal with large amounts of XML information.

XML Path Searching

Path searching is a powerful mechanism for specifying complex search queries in XML documents. This feature is useful for developers who are building XML-based business applications and need to search XML text.

WHAT IS NEW IN ORACLE9/R2

Along with significant investments in improving performance, Oracle9i/R2 contains a number of functional enhancements to Oracle Text. A more detail description of the 9i/R2 features is available in the “Oracle Text Features Overview”.

• Parallel indexing and parallel synchronization for improving index creation and maintenance.
• Users can parallelize contains queries on local partitioned context indexes.
• Extensible lexer where users can take advantage for defining new lexers.
• Classification package for building classification and routing applications.
• DBA tools package for analyzing and tuning text indexes.
• JDeveloper9i Text wizard for generating search applications (figure 12).

• Web query parser. Users don’t need to write query parsers for handling misspelling correction, operators, etc.

• Many performance improvements.

---

**SUMMARY**

Text – in all its varied forms – represents a vast and rich information resource for organizations of all sizes. With Oracle Text, Oracle provides a new set of technologies for easily and securely managing all the enterprises’ structured and text information.

Oracle Text enables application developers to transparently include powerful text searching capabilities into their SQL-based applications. Oracle Text is a core component of other Oracle products like Oracle9iAS Portal, Oracle eBusiness Suite, Oracle Ultra Search and the Oracle Internet File System, as well as the underlying search technology for Oracle.com.

Oracle Text:

• Is totally integrated with the Oracle database.

• Provides the best internationalization features, including multi-lingual search capabilities.
• Exhibits award-winning search quality.
• Helps organizations leverage existing expertise with the Oracle infrastructure.
• Is highly scalable, reliable, and available.

FURTHER READING

3. Oracle Text Features Overview.
7. Oracle Text Discussion Forum (http://otn.oracle.com/forums/text.html/)