



# Veritas Enterprise Vault™ 6.0 Technical Overview: Indexing and Search

A Technical Overview of Indexing and Search in  
Veritas Enterprise Vault 6.0

*Now from Symantec™*



# Veritas Enterprise Vault™ 6.0 Technical Overview: Indexing and Search

## A Technical Overview of Indexing and Search in Veritas Enterprise Vault 6.0

### Contents

<b>Introduction</b> .....	6
<b>Enterprise Vault and AltaVista indexing engine</b> .....	6
<b>Enterprise Vault and Stellent Converters</b> .....	7
<b>Vault Structure and Indexing</b> .....	7
AltaVista Indexing .....	7
Indexing and Enterprise Vault .....	8
<b>Indexing Service</b> .....	10
An Index Per User and Roll-Over Index Volumes .....	10
<b>Indexing administration</b> .....	12
Level of indexing and size of indices .....	12
Backup of the index .....	14
<b>Re-indexing</b> .....	15
Re-indexing to change the index level .....	15
Re-indexing in response to the Indexing Service being flooded .....	15
Re-indexing in response to index corruption or disaster recovery .....	17
<b>Why 1:1 mapping between users and their index is important</b> .....	17
<b>Servicing indexing requests from the user base</b> .....	18
Standard user search applications .....	18
Archive Explorer and Searching None Email Content. ....	20
<b>Advanced search and the Shopping Service</b> .....	21
Advanced searching: Boolean Logic .....	22
Business Accelerators—search persistence .....	22

**Contents** *(cont'd)*

<b>Language support With Enterprise Vault indexing</b> .....	24
<b>Support for encrypted files With Enterprise Vault</b> .....	25
<b>Third-party integration with the Enterprise Vault Software Developer's Kit</b> .....	26
Benefits summary of Enterprise Vault full text indexing services .....	26
<b>Appendix 1: Boolean search logic</b> .....	<b>28</b>
Simple search .....	28
Advanced search .....	29
<b>Appendix 2: Stellent-supported conversion formats</b> .....	<b>30</b>

## **Introduction**

With archives commonly growing to millions or even billions of objects, customers are beginning to realize that a scalable approach to indexing is a critical part of any archiving decision. At the same time, with the increased focus on investigation and legal e-discovery against archived items, companies are carefully scrutinizing the search capabilities of archiving tools versus their requirements. Finally, end users, with the advent of Web search engines (e.g., Google or MSN) and desktop search engines (e.g., MSN Desktop or Google Desktop), are becoming much more accustomed to the search paradigm for finding content and are demanding fast, easy-to-use tools for locating content—whether it be live or archived.

This paper will show how Enterprise Vault has taken an industry-standard indexing technology and modified it to produce a massively scalable, dependable, and cost-effective indexing system, while enabling the end user and investigative search functionality demanded by customers.

## **Enterprise Vault and AltaVista indexing engine**

Enterprise Vault is the leading unstructured content archiving product available in the market today, but this market position was not achieved overnight. KVS first shipped Enterprise Vault in 1999, but its history goes back further, having been initially developed at Digital Equipment Corporation. At the time, Digital also was responsible for one of the most advanced and industry-proven search and indexing engines, AltaVista.

Because of this relationship, the Enterprise Vault developers were able to embed the code of the AltaVista indexing engine into the product, rather than calling a separate executable or service. This means that not only is Enterprise Vault able to offer significant performance benefits, but also the way AltaVista operates and is managed can be tuned to the needs of an archiving application.

The interaction of Enterprise Vault and the AltaVista indexing engine will be the main focus of this paper, and we will explain how Enterprise Vault manages the entire indexing process to give a truly enterprise-class archiving platform.

## Enterprise Vault and Stellent converters

Before we can index an item, we have to understand how to open, or “read,” the document. We have all seen documents on our personal computers that have unassociated extensions, and thus no application is tasked with opening them. If an archiving system encounters an unknown extension, it will be able to archive the document (i.e., place it in the Archive), but it may not be able to index the document and make it available via search applications.

Because of the complexities of understanding the huge number of document types in an average organization and the constant change involved in this task, Enterprise Vault integrates another industry-leading application suite, the Stellent Outside In® document conversion libraries, to convert the documents into a standard format that Enterprise Vault can index. Using this system Enterprise Vault can index approximately 300 different file types. Appendix 2 has a list of the supported file types, and the most up-to-date list of supported file types can be obtained from Stellent.<sup>1</sup>

As will be shown later, during the indexing flow of control, these libraries and the index process are both managed by Enterprise Vault and work together in a seamless fashion to facilitate both the rapid indexing of items and the speedy recall of content.

## Vault structure and indexing

### AltaVista indexing

How the AltaVista indexing engine works is not the subject of this paper; we are instead focused on how Enterprise Vault manages AltaVista. There are, however, a few key basic indexing concepts that first need to be explained.

AltaVista essentially reads the documents that are passed to it, looks at every word within the document, and adds each word to a **word list**, often called an **inverted word list**. This word list will contain the lists of words and the documents in which they appeared. This pairing is called a **unique word location**.

However, not every word located is added to the word list. Certain words are used very often and add little to a search and, more importantly, create an unnecessarily large index file. Words such as “THE” or “AND” are called stop words and are not added to the word list. Many of these “words” can be used in searching, but they will be used in the **Boolean** or **advanced searches** discussed later in the document.

<sup>1</sup> www.stellent.com

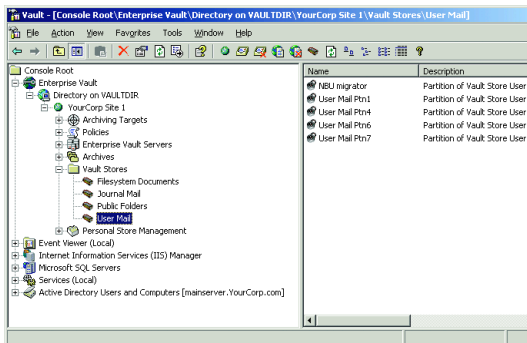
This word list is then broken into a series of files on the storage device. These files are optimized in such a way that when searching, only a portion of the word list needs to be opened and not the entire word list, meaning search results can be obtained faster.

Other improvements in speed come from the concept of preloading the index. The first query of an index will be the slowest as the file on the disk is read into memory first, then the search performed. The Indexing Service will then check to see if the file in memory has the same version number as the file on disk, and if it is the same, it will use the version in memory.

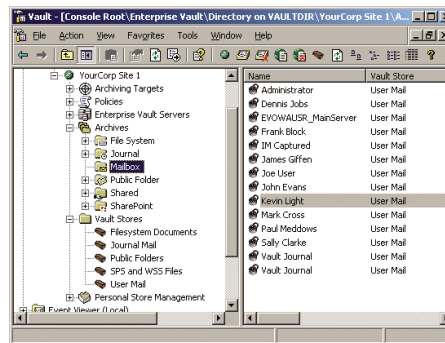
## Indexing and Enterprise Vault

Now that we have seen some of the basic concepts of indexing engine, we can now examine how AltaVista fits into Enterprise Vault.

The technical paper on the Enterprise Vault interaction with storage systems provides more details about the structure of Enterprise Vault, but at a high level we can see that there are four main container structures that make up Enterprise Vault. These can be seen in the Vault Administration Console (VAC) shown in Figures 1 and 2.



**Figure 1.** The VAC shows the Vault Stores on the left. Each Vault Store can contain more than one Vault Store Partition.



**Figure 2.** Archives represent management of logically connected items.

- **Vault Store.** The Vault Store is a virtual representation of content held within Enterprise Vault. The Enterprise Vault administrator groups these items in such a way that it will aid management of Enterprise Vault. Figure 1 shows the Vault Store and Vault Store Partition. The Vault Store is not an indication of what is physically stored on a disk system controlled by Enterprise Vault. A Vault Store contains a series of Vault Store Partitions.

- **Vault Store Partitions.** The Vault Store Partitions are where the actual archived content is stored. This is covered in the white paper on storage.
- **DVS files and collected DVS files.** A Vault Store Partition contains the actual archived content, but this content will reside in one of two main structures. First, when items are originally archived, they are stored inside a DVS file (Digital Vault SaveSet). These files are both compressed and also allow Enterprise Vault to perform **Single Instance Storage**, where identical items are only stored once. So while a DVS file is a single file, this single item may be shared by many users. The second storage structure is the collected file format. Storing items as single DVS files is very fast for access, but as the archive grows, problems will be encountered in storing many millions of separate files. One of the first problems to be seen will be the effects on backup. To aid this, as items age they will be gathered together in “collections” that are more backup-friendly. Access to the files contained within the collections will be slower, but as they are older items on average, they will be accessed less frequently.
- **Archives.** The Archives tab in the VAC (Figure 2) shows that there is a difference between physical storage and management view. The Archive shows the logical view of a collection of items from the same location, typically a single mailbox. The most important aspect of the Archive view is that every archive (e.g., mailbox) will have its own index, and as will be shown later, this is important for scalability. This is a management view, so no content is held in this structure and it does not directly equate to any storage on disk (e.g., Single Instance Storage can happen across Archives).

As hinted above, consideration is required when determining where the index files should be located. If you have chosen to use a WORM storage device or an offline storage system for the archival storage, then this will not be suitable for storing the index file for several reasons. First, as more items are added to a Vault Store, more items are added to the index, and if this is a read-only storage device, then you will not be able to update the index files accordingly. If it is achievable, it will probably be with significant performance reduction. Second, if a user were to perform a search, it would be expected that the index would be quick to respond. If this were an offline storage system, then the index response would be very slow, if it worked at all.

Even if the decision has been made to store the archived items on a WORM or offline media system, the index files would still have to reside on a fast, online disk system. As part of the initial design phase, the storage system needed to house the index files must be determined.



## Indexing Service

The Indexing Service is responsible for both sides of the indexing operation. First, it is the Indexing Service that handles the indexing of items as they pass through to the Storage Service, adding the content to the index. Second, the Indexing Service handles requests from users or applications that are searching the Archive.

Each Enterprise Vault server has a single, dedicated Indexing Service. However, there could be more than one Enterprise Vault server per site, and each of these servers could be configured for a different task. So one server could be responsible for servicing user requests while another is responsible for client requests, allowing the indexing work to be load-balanced across a number of servers.

The actual act of adding content to the index is broken into two main actions:

1. The Stellent converters will create HTML copies of content so that the indexing engine will be able to index this content. As a side effect, this HTML copy is stored for “future proofing” purposes, so that end users or administrators can view archived items even if the original client application for the file type (e.g., Microsoft® Word 95) is no longer available.
2. The indexing engine itself then receives an HTML copy of this content and performs the indexing. Once the item is stored in the index it will be given an ID, placed inside the SaveSet, and passed to the Storage Service.

Once a request is made against the index, the index ID can be referenced against the SaveSet ID in the SQL database, and thus actual items can be recalled from index searches.

## An index per user and roll-over index volumes

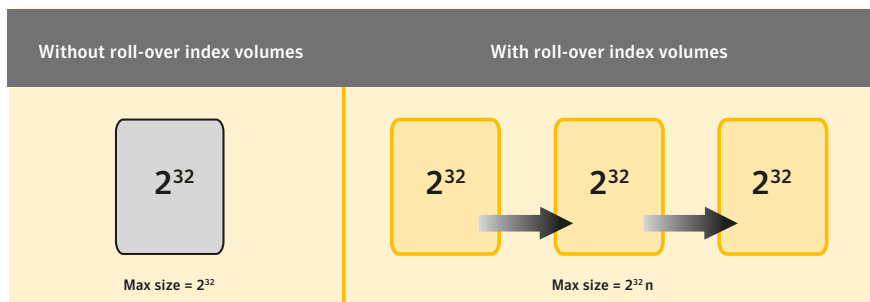
The most critical aspect in determining the upward scalability of an archive is going to be the index, as this is likely to be the element of the system that reaches a “saturation point” long before the file structures or storage devices. Like nearly every file format or storage system, AltaVista index files have an optimal size and a fixed upper limit. AltaVista can handle  $2^{32}$  unique word locations in a single index. The amount of archived data that this equates to is hard to determine, but it does not mean that an index is limited to archiving documents containing  $2^{32}$  words. Once all the “stop words” and repeated words are taken into account, the total amount of words that this equates to is a far greater number, and a single index file can typically reference many millions of emails.

To improve the scalability of the Enterprise Vault index, the first major point already mentioned is that every Archive has its own index. Note that Enterprise Vault creates a separate logical archive for each end-user mailbox archived through mailbox archiving, while maintaining a single archive for archived journal content. Similarly, Public Folder, File System, and SharePoint® content typically reside in separate archives. Finally, it is relevant to mention that Single Instance Storage is preserved across archives within the same Vault Store Partition (as described in more detail in the Enterprise Vault storage white paper).

If all user mailbox archives shared the same index, then clearly it would cause issues in backup but also the index could only scale to  $2^{32}$  items. However, with an index per user, we can see that we are able to scale to  $U \cdot 2^{32}$  items (where U is the number of users).

However, there is still the problem that some users have very large archives and therefore very large indices. To completely alleviate this problem and improve scalability further, Enterprise Vault 6.0 has introduced the concept of **Roll-Over Index Volumes**<sup>2</sup>.

Enterprise Vault 6.0, based on best practice, understands the optimal size that an index should reach. This is below the maximum limit of  $2^{32}$ , but as with the collection of DVS files, the number of files should be minimized to improve search performance.<sup>3</sup> When we pass this practical limit of the index, Enterprise Vault index management automatically creates a new index file and marks the old file as read-only. This means that without Roll-Over Index Volumes, the practical upper limit for a user with a single index file would be  $2^{32}$  unique word locations. As shown in Figure 3, with Enterprise Vault 6.0 this limit is  $2^{32} \cdot n$  (n is the number of roll-overs), meaning that the scalability of the index has been massively, nearly infinitely, improved.



**Figure 3. Roll-Over Index Volumes add even more scalability to the indexing engine over the index per user.**

<sup>2</sup> Enterprise Vault Version 6.0 SP1

<sup>3</sup> See the technical paper on storage for an in-depth discussion on collection and migration.

More time will be spent later in this paper discussing how users of the system interact with the Indexing Services, but Roll-Over Index Volumes are a good starting point to state that indexing is not just about how we intelligently put content into the index, but also how users interact with the Indexing Services. If we have created the concept of a single-user archive containing many index files, then how does the end user know which of these index files to search? The simple answer to this is that they do not need to. Enterprise Vault uses the concept of search federation to ensure that when a user searches an archive, all indices are searched and the end user is presented with a consolidated set of results.<sup>4</sup> The end users do not know that several searches were needed to service their request.

Roll-Over Index Volumes are the best example of how Enterprise Vault has taken a proven indexing system and modified it for use in an archiving system. This is only possible because Enterprise Vault has intimate knowledge of the AltaVista indexing system. This gives Roll-Over Index Volumes near limitless scalability and allows more information to be stored in a single Archive, meaning reduced overall management costs.

### Indexing administration

The following section will detail the important administration and management tasks that can be performed on the indices within Enterprise Vault. It's important to note that not all of these tasks are manual, and many, like Roll-Over Index Volumes, are carried out automatically by the Indexing Services.

### Level of indexing and size of indices

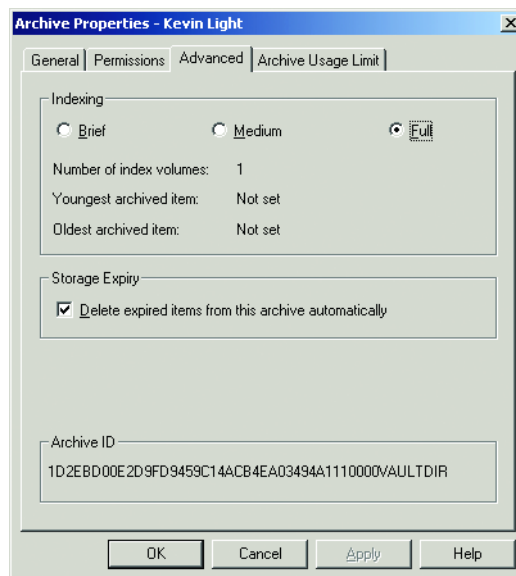
There are various “levels” of indexing that can be achieved with Enterprise Vault: **brief**, **medium**, and **full**. The precise definitions of what can be achieved with each is discussed below, but at a simple level, the higher the level of detail held in the index, the greater the depth of analysis that can be performed on the index at a later date.

- **Brief indexing** means that you are only able to perform searches against the metadata of archived items. For example, you could use this indexing level to request emails sent within a certain date range.
- **Medium indexing** captures every keyword and, for example, allows you to search for documents containing the word “merger.”

<sup>4</sup> Full search federation with all applications is not delivered in Enterprise Vault 6.0 SP1 but will be included in later service packs.

- **Full indexing** will allow more advanced searching to locate content faster, such as being able to search for entire phrases (rather than individual keywords). For example, you could use this indexing level to search for documents containing the phrase “Acme Corporation merger.”

As can be seen in Figure 4, changing the level of indexing is a very simple administrative task. Once the index level has been changed in the Vault Admin Console, all new content will adopt the new level.<sup>5</sup> There is a Site Level property that can be set to fix the level of indexing for all archives created in that Enterprise Vault Site.



**Figure 4. An index can be at three levels and easily changed**

There is a penalty for increasing the level or depth of indexing: the amount of storage required for the index. There is a less dramatic effect on the speed of the Indexing Service, and this effect is minimized further by using the Replay Index functionality (discussed later in this document). The major penalty is going to be on the size of the index files held on disk.<sup>6</sup> As Figure 5 shows, the higher the level of indexing, the greater the amount of disk space needed to house the index files.

<sup>5</sup> See the later section on re-indexing to see how changing the level of indexing affects content already archived.

<sup>6</sup> When referring to the size of the index, we are referring to the size of all the index files associated with the archive and the Roll-Over Index process.

Index level	Index size as a percentage of total archive size
Brief	3%
Medium	8%
Full	12%

Figure 5. Estimated size of an index relative to the total size of archive

As indicated in Figure 5, the size references are estimates. It is impossible to give exact ratios for the size of the index. The exact size of the index file will be determined by the nature of the content being archived.

### Backing up the index

Another critical administrative duty is backing up the Enterprise Vault indices as they are as important as any other business information in the organization.

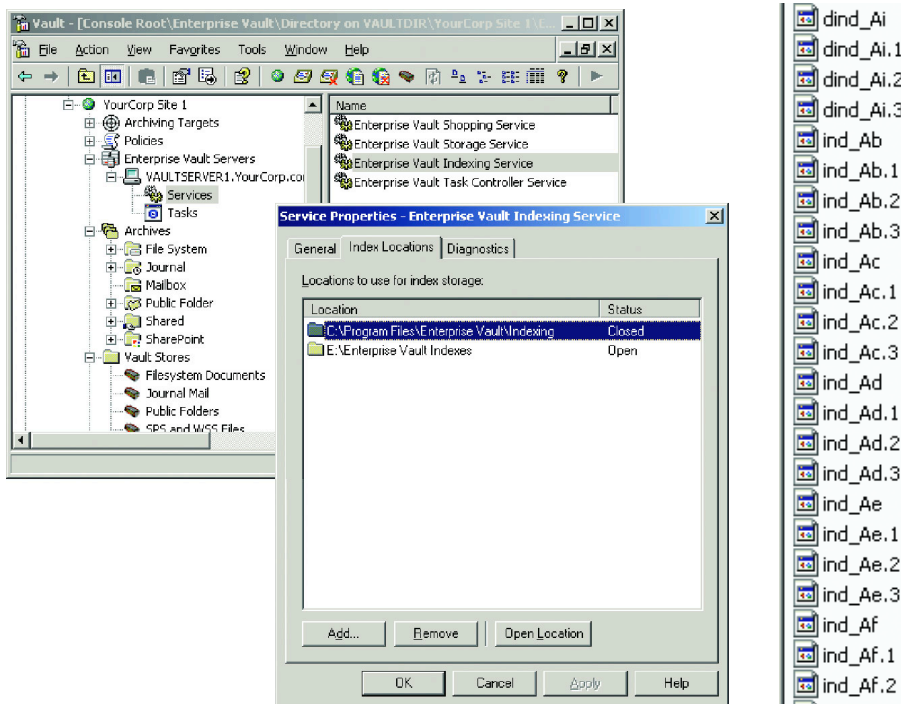


Figure 6. The location of the index files is controlled by the Index Service, and the index files on disk are split and optimized for rapid access.

As shown in Figure 6, the index files are a series of files on disk. They can, in essence, be treated the same way as any other file as far as backup is concerned, but Enterprise Vault was designed to make backup and restoration of the index files as efficient as possible. The Enterprise Vault Administration Console shows the locations of the index files. More than one location may need to be backed up if the particular server has more than one Vault Store.

If the Archiving services are suspended or placed in backup mode, no new information will be added to Enterprise Vault or the indices. This means that at this point in time the index files are read-only and a backup can easily be performed. A further option is also available where the Indexing Service can be stopped and the files backed up. This will give a full true copy of the index, sometimes considered best in backup terms. However, the downside to this approach is that no one will be able to access the index during the backup.

### **Re-indexing**

#### **Re-indexing to change the index level**

As was shown in Figure 4, the level of indexing for all new content entering Enterprise Vault can be changed, but what about the content already archived? How easy is it to change the level of indexing of this content?

When you change the level of indexing on an archive as above, you will change the index level for all content in that archive. This “re-indexing” is achieved by crawling the physical DVS files associated with that archive, removing all previous content from the index, and then adding new content, with the new indexing level.

#### **Re-indexing in response to the Indexing Service being flooded**

Re-indexing of archived content is a core activity of the Indexing Services of Enterprise Vault and is often referred to as **Replaying the Index** (or **Replay Index**).

Since the primary design goal of Enterprise Vault is data integrity, at no point in the lifecycle of archived content should it be lost or corrupted. If Enterprise Vault determines that the quantity of data being archived is too great for the current throughput abilities of the system (with indexing), it may turn off indexing until the flood of data subsides. If this happens, an alert is sent to the administrator. This automatic throttling of indexing ensures that the queues that service content coming into Enterprise Vault do not grow beyond control and content is always held securely in the archive.

During this time, the Archiving Service concentrates solely on ensuring that content is correctly written to disk and, at the same time, keeps track of content that has not been indexed. The Indexing Service itself is then running in Replay mode, meaning that content already stored in the archive is being used to create the index as opposed to content en route to the archive.

As shown in Figure 7, determining if the system has been placed in Replay mode is very important to the flow of control of indexing. If the system is in Replay mode, items will take a different route as they are committed to the archive, as opposed to when Replay Index is not in operation. Regardless, once an item is called for indexing, the same functions will apply—namely conversion, then index.

Clearly, replaying the index needs to be considered carefully and should only be used when the Enterprise Vault services themselves become the bottleneck. Putting Enterprise Vault into Replay Index mode because of a bottleneck in the storage system will not help as this will, in effect, increase storage throughput (data has to be reread from the archive). In general it is very rare that the Enterprise Vault servers are the bottlenecks, as a correctly designed and provisioned Enterprise Vault server will have faster performance than the attached storage.

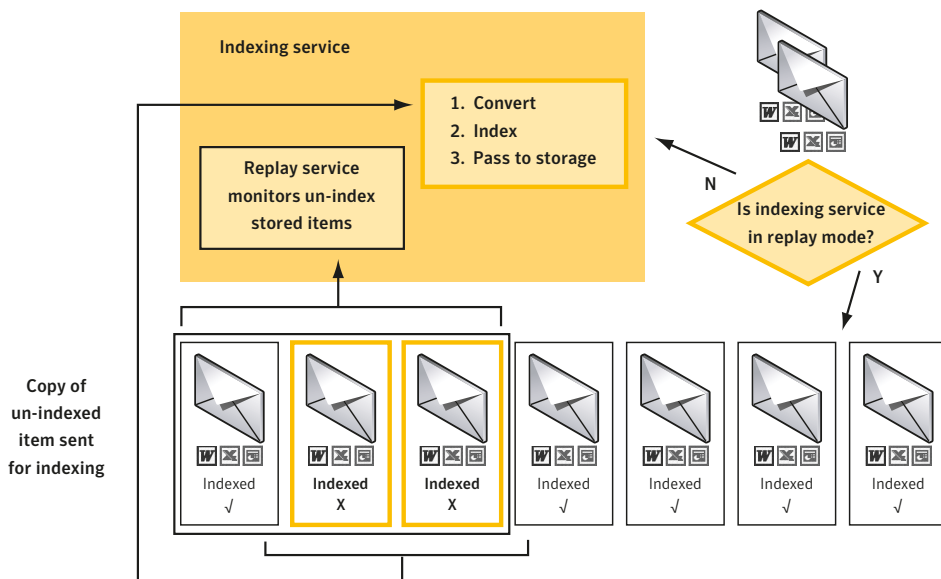


Figure 7. Before items are sent to the Indexing Service, determine if the system is in Replay mode.

### **Re-indexing in response to index corruption or disaster recovery**

As previously mentioned, avoiding data corruption is the number one design goal of Enterprise Vault, but that does not mean that this data corruption or system failure can be ignored. Replay Indexing functionality is vital to recover an Enterprise Vault installation from a system failure.

Imagine the following scenario:

- The machine on which Enterprise Vault is running physically fails and you discover that your backup of this machine has also failed.
- The indices for Enterprise Vault were held on direct-attached disks on the Enterprise Vault Server and have been lost.
- The physical archive files were held on a server other than Enterprise Vault and were recoverable.
- Replication within Microsoft SQL Server and the fact that Enterprise Vault directory database is held within this system means that the directory database, vital for the operation of Enterprise Vault, is still operational.

What should be mentioned first is that there is no better alternative to a good and tested backup, but in the scenario above, the Replay Indexing facilities of Enterprise Vault means that the index could be completely recovered from the underlying archived data. The Indexing Service would be placed in Replay Index mode and notified that the entire archive has not been indexed and so it would re-create the entire index.

In addition, since the indices can easily be backed up, it is far less likely that the backups would fail and that the above situation would occur in the first place.

### **Why 1:1 mapping between users and their index is important**

One of the reasons why we want an index per end-user mailbox archive is shown when considering re-indexing.

Imagine you have an e-discovery request focused on 10 out of 10,000 users. You may decide that you want to change the level of indexing for these users, and since each user has their own index volume, this is very easy to achieve. If the archiving system had an index design where all users were in the same index file, then this would mean that 10,000 users would have to be re-indexed, and not just the 10 that are involved with the request.



Similarly, if specific indices are corrupted (which will happen with all indexing systems, given that hard disks fail from time to time), the federated nature of Enterprise Vault indices allows for only the specific, affected indices to have to be rebuilt, versus requiring a rebuild of all of the archive indices. This enables much faster recovery from failure as well as greater overall system resiliency.

### **Servicing indexing requests from the user base**

So we have now shown the functions and configurable options available to add content to the index, but as previously stated, archiving is not about putting content into the “box.” It’s about being able to recover content in a timely manner when it is required. Clearly having a full text index makes finding the required content many times faster.

The next section will consider some of the built-in search applications delivered with Enterprise Vault that will aid the end user as well as the business as a whole.

### **Standard user search applications**

One of the key advantages of archiving is that the end user can easily locate information by using simple keyword searches. Previously we stated that each Enterprise Vault server can have its own Indexing Service to carry out all interactions with the index. We will now look at some of the applications that utilize those services to enable user-based searching.

As part of the Microsoft Outlook® plug-in installed with Enterprise Vault, administrators are able to give users access to the simple search within Outlook. In fact, the search application is driven by an HTML page installed on the Enterprise Vault server. Every Enterprise Vault server needs to have Microsoft Internet Information Server installed, as all requests for content from the vault are via HTTP. Also installed as part of the standard installation is searchO2k.asp. This Web page is the simple search page that is viewed once a user clicks on the search button in Outlook. Figure 8 shows that users do not have to guess how to interact with searchO2k.asp. This search page is embedded into Outlook and is considered part of Outlook. This means users are able to utilize the search with little or no training and increase their productivity considerably.

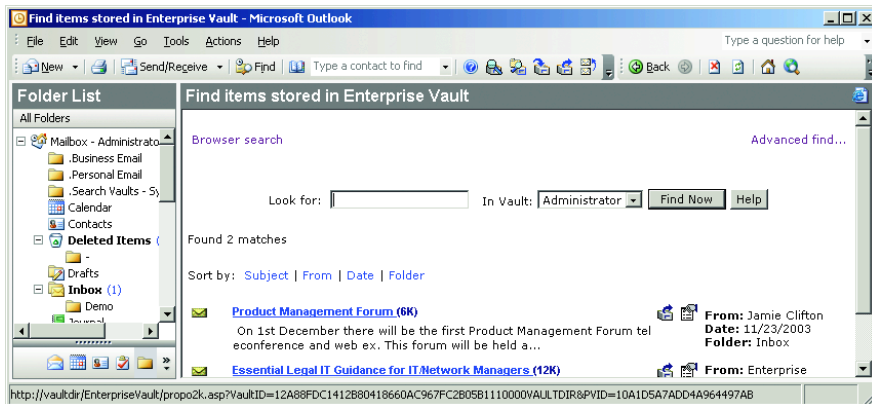


Figure 8. The basic search page given to users in Outlook shows matches in the message and attachments.

The page search02k.asp interacts directly with the Indexing Service and renders the completed search pages to send back to the end user. The result indicates to the user whether the content was located in the email message body or attachment. Users can also click and open the item or restore the item back to the original location.

It is important to note that users searching (or using any of the search applications) can only view content that they have rights over. They are not presented with any items in the list over which they do not have “read” access rights.

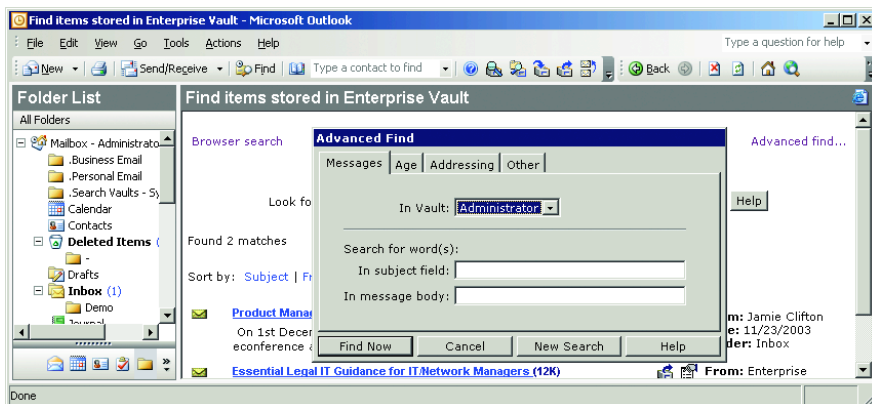


Figure 9. Advanced find allows refining of the search criteria to allow more rapid location of content.

SearchO2k.asp also gives users the option to refine the search criteria using “advanced find.” A simple search only allows users to choose a key phrase, but with advanced find, users can refine the search by placing time limits or TO:/FROM: information and locate the content faster. Again, end users are presented with this extended functionality in a convenient manner. Figure 9 shows that the advanced search options are as easily available in the standard search options and appear as part of Outlook.

## Archive Explorer And Searching None Email Content

The search applications above are both email-specific, but searches within Enterprise Vault are not limited to email information, as all content archived into Enterprise Vault is indexed and available via search. Exactly the same search is built into the Archive Explorer application (Figure 10), which allows users to search all content in the archive, for example, files and emails. Archive Explorer offers additional functionality by allowing users to see a folder “hierarchy” representing the folders (in email or in the file system) from which the items were archived (even if shortcuts are no longer resident in those folders).

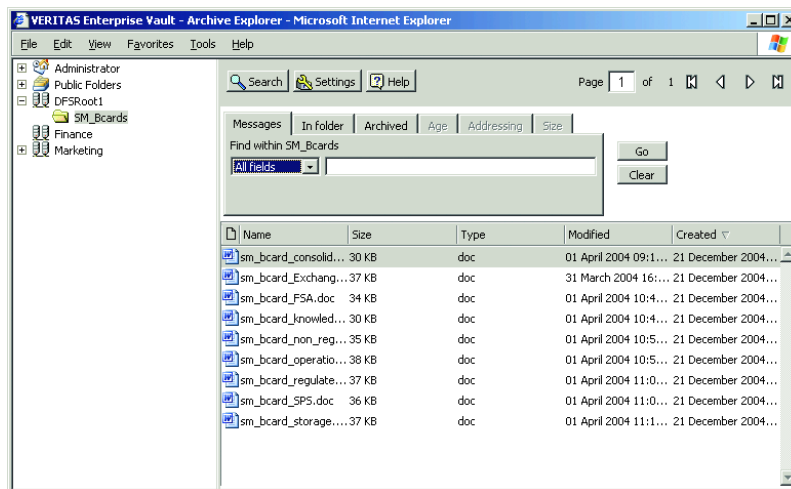
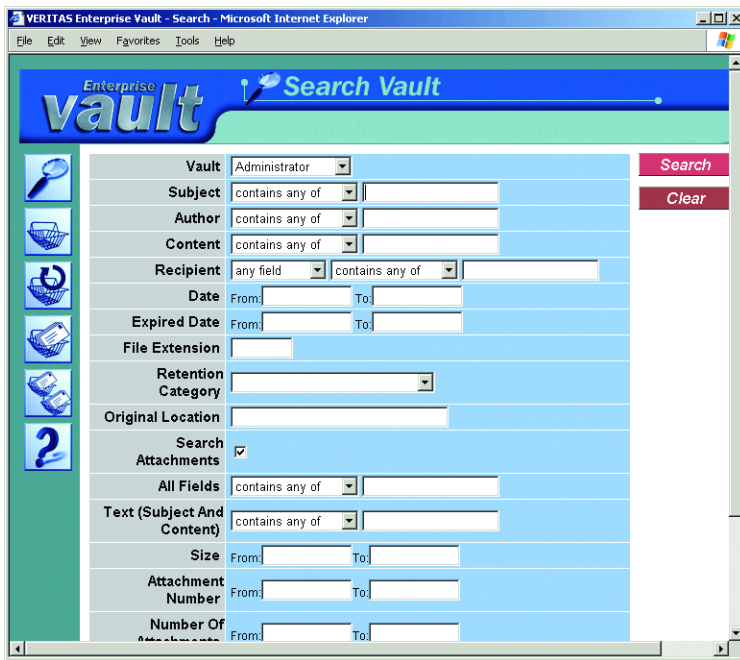


Figure 10. Archive Explorer also allows access to the Indexing Service.

In addition to having the same search abilities as the Outlook search add-in, Archive Explorer is also delivered to users as a Web page, allowing end users to search and access archived content without any client add-ins.

## Advanced search and the Shopping Service

Another search application available to end users is search.asp (Figure 11). In addition to using the Indexing Service, this search application uses the **Enterprise Vault Shopping Service**. This search page is again delivered to end users via a Web page, and like the advanced options in search02k.asp, the application accepts not only the keyword search but also parameters to refine the search, for example, date ranges.



**Figure 11. The browser search allows users to interact with the Indexing Service and the Shopping Service. This search allows users to save baskets of search results but it also allows them to bulk restore items.**

The search here has other features for the end user: the **shopping basket** and **bulk restore**. The Shopping Service is an Enterprise Vault feature allows the end user to create “baskets” of searches. Once search results are listed, the user has the option to create a basket of these searches with their associated criteria and return to them at a later date. Note: With all the other user searches, once the user closes the search page, the search is lost; but with the Shopping Service these results are retained in a basket for future review.

In addition, browser search also allows the user to bulk restore items to their original location. The simple search in Outlook has the option to allow users to restore items on an item-

by-item basis. The Shopping Service allows all items listed in the search to be returned to their original location with the click of a button. This is a very useful administrative feature that will, for example, allow the easy recovery of a large number of items from a user's archive back to their mailbox.

### **Advanced searching: Boolean logic**

All of the searches listed above are based on simple keyword matching, and there are additional advanced features to refine the search—for example, by date or by sender. Searches can also incorporate Boolean logic into the search string to give increased flexibility.

We have already mentioned that AltaVista has the concept of stop words, or words that are not indexed, and these stop words are useful in allowing more advanced searching. For example the word AND can be used to search for “this” *and* “that” (meaning a document has to have both words).

Details of the Boolean operators and how they apply to each search interface can be found in Appendix 1.

### **Business Accelerators—search persistence**

The search applications that have already been outlined are examples of how users would want to interact with the Indexing Service. When an organization or business needs to search the archive, it has very different goals from those of an end user. For example, many organizations are looking to email archiving to automate the process of electronic discovery, or e-discovery, where email and other electronic content needs to be searched as a part of a legal investigation. A common scenario would involve an organization looking for all email between a listed set of individuals over a given date range containing a set of keywords. Because of this, Enterprise Vault approaches an enterprise-wide (as opposed to end-user) search uniquely:

1. The security of the user applications described previously is designed to limit the amount of information that a user can access, to ensure that they are only able to see the information that they have rights to. When an organization wants to perform enterprise-wide searches (e.g., for e-discovery), it needs access rights over all indices.
2. Not all users should have the same rights when interacting with the Index Service. For example, when interacting with the index, we may want to declare that some technical people can create

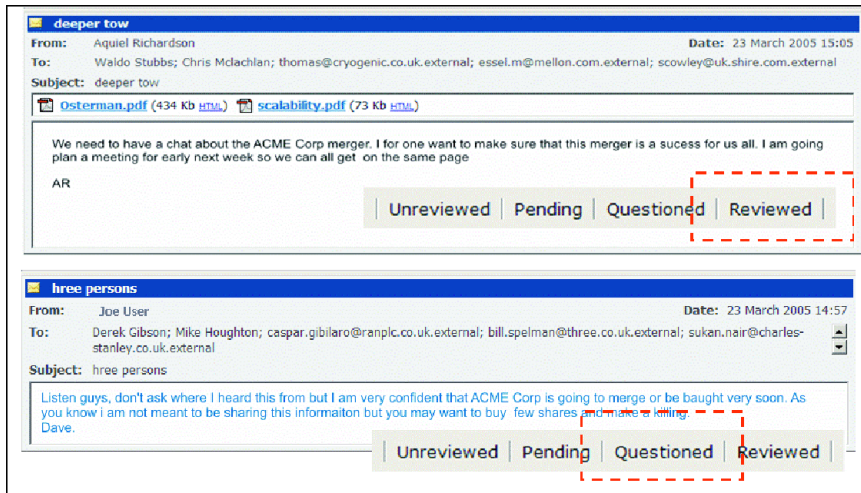
searches but they cannot view the results of the search. In essence, we should be able to create roles that are associated with the search process.

3. With the exception of the Shopping Service, when a user closes the search screen, the search results are lost. The size of the search results when searching the entire archive can be huge, so it's unlikely that a person or team of people will be able to review all the search results in one sitting. There must be some persistence of the search, so that the review team can revisit the search.
4. If an item is determined as "always" belonging to a certain category, this knowledge needs to be retained. For example, if an initial search defines an item as spam, and your corporate policies have determined that this will always be the case, then this particular item can be excluded from subsequent reviews since it has already been categorized. This will dramatically improve the speed of future searches.

Security is comparatively easy to solve, as there is the option to create an administrator account that has rights over the entire archive. As Enterprise Vault is a read-only data source, this account is, in essence, read-only, which means no content can be changed—even by the administrator. There are even fewer security considerations when using Microsoft Exchange Journaling, as in this scenario all emails being transmitted through the system will be delivered to a single mailbox, and this content will then be moved to Enterprise Vault quickly. Since all the messages come from the same mailbox, defining a user that will have effective search rights over all the organizational content is very easy.

Search persistence is the ability to turn off a search application while retaining the search results. This is a more complex matter that is solved by joining the full text indexing abilities of AltaVista and the structured advantages of Microsoft SQL Server. The initial search is performed by AltaVista and the results retained in a SQL Server database.

Enterprise Vault provides Business Accelerators, which are applications developed for the enterprise-wide search scenarios described above. For e-discovery, the application is called Discovery Accelerator. Discovery Accelerator creates cases that can be viewed as e-discovery requests; each of these cases has its own SQL data containing the matches that were located by AltaVista. This data can then be reviewed, marked, and potentially exported by a user or group of users over a period of time.



**Figure 12. Discovery Accelerator extends the search index by persisting the search results and allows review teams to determine if content is relevant to the discovery process.**

Figure 12 shows the basic aim of Discovery Accelerator: to securely allow a team of reviewers to categorize content in the archive as being important to the case in question. As can be seen in Figure 12, a search for the term “ACME Corp Merger” will return some results that are important to your specific case and some that are not. Groups of reviewers will be able to determine which are important and which are not and mark them appropriately. Indexing is key to this application, as finding the content initially would be nearly impossible without a robust, scalable indexing system.

### Language support with Enterprise Vault indexing

The AltaVista search engine provides the capability to index full Unicode texts using the UTF-8 encoding. Unicode is a multi-byte encoding framework that provides for 2<sup>32</sup> character positions, of which only about 2<sup>16</sup> are filled to date.

Unicode is aimed at multilingual environments and internationalization. AltaVista has limited support for automatic parsing of texts into words for the common European character sets.

Since the index is in Unicode, searching can be language-specific. For example, a search for “éléphant” would only yield the French variant of the word (more specifically the accented “é” in the word, regardless of the language it was written in).

### **Support for encrypted files with Enterprise Vault**

Many organizations are beginning to explore encryption solutions for email and file content—particularly for email being sent to external parties. Organizations looking at a long-term retention strategy for electronic content need to think about the long-term search-ability for this content as well.

By default, encrypted content will be archived by Enterprise Vault. The content properties themselves (e.g., sender, recipient, subject, etc., for email or file name, date, etc., for files) are not encrypted and thus are indexed as normal by Enterprise Vault. Users can therefore search for content that is encrypted by any of the normal properties (or metadata) of email messages or file content. Furthermore, end users who have access to the appropriate decryption mechanism for the content desired (e.g., the decryption keys) can retrieve and decrypt the required content as with normal files or messages.

However, encryption presents a challenge for organizations, as “by default” encrypted content (messages, attachments, files, etc.) cannot be opened by Enterprise Vault and therefore cannot be indexed. Thus, while you could search all messages (including encrypted messages) for keywords in the subject line, you cannot by default search encrypted content itself. For organizations that need to be able to search all content, this creates an important decision point.

As a simple step, Enterprise Vault logs every encrypted item as it is archived, allowing an organization to identify encrypted items and go back to the creator (e.g., sender of the email) to ensure that they are using a company-approved encryption system.

In addition, Enterprise Vault was designed with this problem in mind and provides a custom filter API that allows third-party applications (e.g., encryption/decryption engines) to view and modify content prior to it being indexed and archived. Through this, encryption providers can build plug-ins to automatically decrypt content prior to Enterprise Vault archiving the content.

As an example, Entrust Software has built an integration using this API to allow any PKI-based encryption system to be integrated with Enterprise Vault and have content decrypted prior to archiving.[What is “this API”? Is this a reference to the custom filter API mentioned above? If so, perhaps we should say “the Enterprise Vault API” or “this custom filter API” instead of “this API.”] As a second benefit beyond searching, this ensures that archived content itself can be retrieved and viewed even if the keys are no longer available.



### Third-party integration with the Enterprise Vault Software Developer's Kit

The aim of Enterprise Vault is to be the enterprise archiving platform of choice, meaning that many different types of content can be stored in the Enterprise Vault and that this content can be accessed and exploited in many different ways. To facilitate this, Enterprise Vault has a full Software Developer's Kit (SDK) built around it, so organizations can add custom data to Enterprise Vault and then utilize this content in whatever way they require.

The Search API allows developers to create simple searches, and it returns references to the items that met the search results. The items themselves are not returned via this API, only the reference to the item; other APIs can be used to retrieve the items. This API allows a company to include Enterprise Vault search results as a part of a larger portal or Enterprise Content Management search framework.

Furthermore, the Enterprise Vault Custom Properties API allows additional properties to be added to items at the time of archive. These properties can then be searched by applications such as Discovery Accelerator. For example, a customer could choose to automatically look for emails with the keywords "For Internal Use Only" in the subject line and then "tag" those items with an index property called "Confidential." Users could then use Discovery Accelerator to search only on messages with the "Confidential" tag.

For more information on these and other APIs, see the Enterprise Vault API documentation.

### Benefits summary of Enterprise Vault full text indexing services

1. **Integrated indexing.** Indexing is built into the heart of Enterprise Vault and is not an add-on or an executable that is called, leading to a loss of efficiency.
2. **Flexible document conversion.** The use of Stellent document converters enables a huge number of document types to be fully indexed. As new document types become available, they are likely to be supported by Enterprise Vault.
3. **Ease of management.** Enterprise Vault creates a virtual management view of the Indexing Services and objects in the Vault Admin Console to aid administration. This virtual view of the indexing and the Indexing Services means that the administrator needs to visit only one location to control indexing and need not be concerned with the physical data on the disk, except to consider backup.

4. **Federated indices.** Enterprise Vault provides greater scalability by having a large number of smaller indices (with one or more per archive), while automatically rolling over indices based upon data volume. This provides improved performance and a faster rebuild process if index corruption happens.
5. **Re-indexing as a core function.** It's very easy with Enterprise Vault to keep indexing at a low level and the index small. Re-indexing on demand enables in-depth examination of specific areas of the archive. Furthermore as there is an index per user, not everyone needs to be re-indexed—only the relevant users.
6. **Time-proven.** The AltaVista engine itself is a well-tested and proven indexing solution. In addition, Enterprise Vault has shown clear leadership in using this indexing engine in an innovative manner to create the market-leading archiving application.

## Appendix 1: Boolean search logic

Below is a summary of how a simple search with Enterprise Vault user search applications can be refined using Boolean logic.

### Simple search

The simple search is available through Outlook and Archive Explorer, and the following operators can be used.

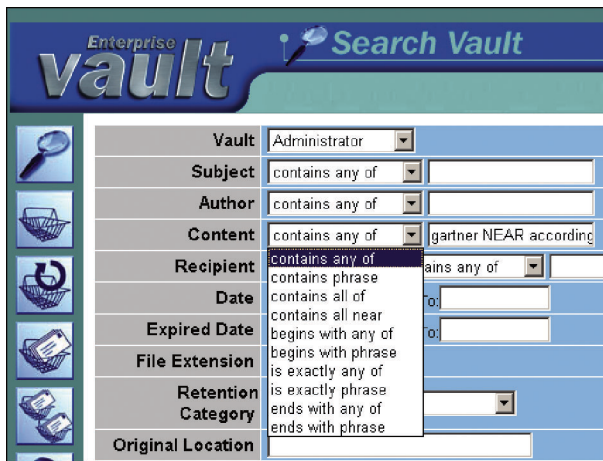
OR	Entering terms with only a between them
AND / +	Placing a + before both the terms that in Advanced Search are linked by AND
NOT / -	Preceding a term with - ensures that items do not contain the search term
“”	Search for the exact phrase between the quotes
*	Wildcard searching with unlimited characters in between Kev* finds Kevin Kev*n finds Kevin
?	Wildcard searching but only allows for one character missing Kevi? Finds Kevin Kev? Does not find Kevin

Note: There is no “NEAR” operator in the simple search.

Note: The simple search is not case sensitive.

## Advanced search

The advanced search offers additional features over and above Boolean logic.



The screenshot displays the 'Enterprise vault Search Vault' interface. On the left, there is a vertical navigation menu with icons for search, basket, refresh, and other functions. The main area contains a search form with the following fields and options:

Field	Options/Value
Vault	Administrator
Subject	contains any of
Author	contains any of
Content	contains any of gartner NEAR according
Recipient	contains any of, contains phrase, contains all of
Date	contains all near, 0:
Expired Date	contains all near, 0:
File Extension	begins with any of, begins with phrase
Retention Category	is exactly any of, is exactly phrase, ends with any of, ends with phrase
Original Location	

Figure 13. Much Boolean logic is built into the advanced search interface.

Most important is the ability to refine the search via a date range through simple fields in the search form. In addition, many of the constructs are built in, and the NEAR operator can be used in this interface.

## Appendix 2: Stellent-supported conversion formats

Below is a list of all the document types that are supported by the Stellent conversion engine. In effect, this is a list of all the documents Enterprise Vault can index. There are frequent updates to this list, so for more details go to [www.stellent.com](http://www.stellent.com) ([www.stellent.com/stellent3/groups/mkt/documents/nativepage/outside\\_in\\_supported\\_platforms.pdf](http://www.stellent.com/stellent3/groups/mkt/documents/nativepage/outside_in_supported_platforms.pdf)).

Word Processor Formats	
Word Processors—Generic Text	
ANSI Text (7- and 8-bit)	ASCII Text (7- and 8-bit)
EBCDIC HTML through 3.0 (some limitations)	IBM (FFT, all versions)
IBM Revisable Form Text (all versions)	Microsoft Rich Text Format (RTF, all versions)
Text Mail MIME (no specific version)	Unicode Text (all versions)
WML (Version 5.2)	
Word Processors—DOS Word Processors	
DEC WPS Plus (WPL, through 4.1)	DisplayWrite 2 and 3 (TXT, all versions)
DisplayWrite 4 and 5 (through Release 2.0)	Enable (3.0, 4.0, and 4.5)
First Choice (through 3.0)	Framework (3.0)
Lotus Manuscript (Version 2.0)	IBM Writing Assistant. (1.01)
MASS11 (versions through 8.0)	Microsoft Word (versions through 6.0)
Microsoft Works (versions through 2.0)	MultiMate (versions through 4.0)
Navy DIF (all versions)	Nota Bene (Version 3.0)
Novell WordPerfect (versions through 6.1)	Office Writer (Versions 4.0–6.0)
PC-File Letter (versions through 5.0)	PFS:Write (Versions A, B, and C)
Professional Write (versions through 2.1)	Samna Word (versions through Samna Word IV+)
SmartWare II (Version 1.02)	Wang PC (IWP, versions through 2.6)
WordMARC (versions through Composer Plus)	XyWrite (versions through III Plus)
WordStar (versions through 7.0)	WordStar 2000 (versions through 3.0)
Q&A (Version 2.0)	Sprint (versions through 1.0)
Total Word (Version 1.2)	Volkswriter 3 and 4 (versions through 1.0)
DEC WPS Plus (DX, through 4.0)	

## Veritas Enterprise Vault 6.0 Technical Overview: Indexing and Search

Word Processors—Windows® Word Processors	
Adobe FrameMaker (MIF, Version 6.0)	Hangul (Version 97 and 2002, text only)
JustSystems Ichitaro (Versions 5.0, 6.0, 8.0–13.0, 2004)	JustWrite (versions through 3.0)
Legacy (versions through 1.1)	Lotus AMI/AMI Professional (versions through 3.1)
Lotus Word Pro (Version 96 through Millennium Edition 9.6, text only)	Microsoft Write (versions through 3.0)
Microsoft Word (versions through 2003)	Microsoft WordPad (all versions)
Microsoft Works (versions through 4.0)	Novell Perfect Works (Version 2.0)
Novell/Corel WordPerfect (versions through 12.0)	Professional Write Plus (Version 1.0)
Q&A Write (Version 3.0)	Star Office/Open Office Writer (Star Office Versions. 5.2, 6.x, and 7.x)
Open Office version 1.1 (text only)	WordStar (Version 1.0)
Word Processors—Macintosh® Word Processors	
MacWrite II (Version 1.1, Versions 1.02–3.0)	Microsoft Works (Mac, versions through 2.0)
Microsoft Word (Mac, Versions 4.0–2004)	Novell WordPerfect (Versions 1.02–3.0)
Spreadsheet Formats	
Enable (Versions 3.0, 4.0, and 4.5)	First Choice (versions through 3.0)
Framework (Version 3.0)	Lotus 1-2-3 (DOS and Windows, versions through 5.0)
Microsoft Multiplan (Version 4.0)	Lotus 1-2-3 (OS/2, versions through 2.0)
Lotus 1-2-3 Charts (DOS and Windows, versions through 5.0)	Lotus 1-2-3 for SmartSuite (Versions 97–Millennium 9.6)
Microsoft Excel Charts (Versions 2.x–7.0)	Lotus Symphony (Versions 1.0, 1.1, and 2.0)
Microsoft Excel (Mac, Versions 3.0–4.0, 98–2004)	Microsoft Excel (Windows, Versions 2.2 through 2003)
Microsoft Works (DOS, versions through 2.0)	Microsoft Works (Mac, versions through 2.0)
Mosaic Twin (Version 2.5)	Novell Perfect (Version 2.0)
PFS: Professional Plan (Version 1.0)	Quattro Pro (DOS, versions through 5.0)
Quattro Pro (Windows, versions through 12.0)	SmartWare II (Version 1.02)
Star Office/Open Office Calc (Star Office Versions 5.2, 6.x, and 7.x)	Open Office version 1.1 (text only) SuperCalc 5 (Version 4.0)    VP Planner 3D (Version 1.0)
Microsoft Works (Windows, versions through 4.0)	

## Veritas Enterprise Vault 6.0 Technical Overview: Indexing and Search

Presentation Formats	
Corel/Novell Presentations (versions through 12.0)	Harvard Graphics for DOS (Versions 2.x and 3.x)
Harvard Graphics(Windows versions)	Freelance (Windows, versions through Millennium 9.6)
Freelance for OS/2 (versions through 2.0)	Microsoft PowerPoint (Windows, Versions 3.0–2003)
Microsoft PowerPoint (Mac, Versions 4.0, 98–2004)	StarOffice/OpenOffice Impress (StarOffice 5.2, 6.x, and 7.x)
OpenOffice 1.1 (text only)	
Graphics Formats	
Adobe Illustrator (versions through 7.0, 9.0)	Adobe Photoshop (PSD, Version 4.0)
Adobe Acrobat (PDF, Versions 2.1, 3.0–6.0, Japanese)	Adobe FrameMaker graphics (FMV) (Vector/raster through 5.0)
AutoCAD Interchange and Native Drawing formats (DXF and DWG)	Ami Draw (SDW) (Ami Draw)
AutoShade Rendering (RND, Version 2.0)	AutoCAD Drawing (Versions 2.5–2.6, 9.0–14.0, 2000i, and 2002)
Bitmap-BMP, RLE, ICO, CUR, OS/2 DIB & WARP (all versions)	Binary Group 3 Fax (all versions)
Corel Clipart format (CMX, Versions 5–6)	CALS Raster (GP4) (Type I and Type II)
Corel Draw (CDR with TIFF header, Versions 2.x–9.x)	Corel Draw (CDR, Versions 3.x–8.x)
Computer Graphics Metafile (CGM) (ANSI, CALS NIST version 3.0)	Encapsulated PostScript (EPS) (TIFF header only)
GEM Paint (IMG, no specific version)	Graphics Environment Mgr (GEM) (Bitmap and vector)
Graphics Interchange Format (GIF, no specific version)	Hewlett Packard Graphics Language (HPGL) (Version 2)
IBM Graphics Data Format (GDF, Version 1.0)	IBM Picture Interchange Format(PIF) (Version 1.0)
Initial Graphics Exchange Spec (IGES, Version 5.1)	JFIF (JPEG not in TIFF format) (all versions)
JPEG (including EXIF) (all versions)	Kodak Flash Pix (FPX) (all versions)
Kodak Photo CD (PCD, Version 1.0)	Lotus PIC (all versions)
Lotus Snapshot (all versions)	Macintosh PICT1 and PICT2 (bitmap only)
MacPaint (PNTG, no specific version)	Micrografx Draw (DRW) (versions through 4.0)
Micrografx Designer (DRW, versions through 3.1)	Micrografx Designer(DSF) (Windows 95, version 6.0)
Novell PerfectWorks (Draw, Version 2.0)	OS/2 PM Metafile (MET) (Version 3.0)
Paint Shop Pro 6 (PSP) (Windows only, Versions 5.0–6.0)	PC Paintbrush (PCX and DCX) (all versions)
Portable Bitmap (PBM, all versions)	Portable Graymap (PGM) (no specific version)
Portable Network Graphics (PNG, Version 1.0)	Portable Pixmap (PPM) (no specific version)

## Veritas Enterprise Vault 6.0 Technical Overview: Indexing and Search

Postscript (PS) (Level II)	Progressive JPEG (no specific version)
Sun Raster (SRS, no specific version)	Star Office/Open Office Draw [Star Office 5.2, 6.x, and 7.x, and OpenOffice version 1.1 (text only)]
TIFF (versions through 6)	TIFF CCITT Group 3 & 4 (versions through 6)
Truevision TGA (TARGA)	Visio (preview) (Version 4)
Visio (Versions 5, 2000–2003)	Windows Enhanced Metafile (EMF) (no specific version)
WBMP (no specific version)	WordPerfect Graphics (WPG and WPG2) (versions through 2.0, 7, and 10)
Windows Metafile (WMF) (no specific version)	X-Windows Bitmap (XBM) (x10 compatible)
X-Windows Dump (XWD) (x10 compatible)	X-Windows Pixmap (XPM) (x10 compatible)
<b>Compressed Formats</b>	
GZIP (all versions UUEncode)	LZA Self Extracting Compress (all versions UNIX Compress)
LZH Compress (all versions UNIX TAR)	Microsoft Binder (Ver 7.0-97 ZIP, PKWARE ver through 2.04g conversion of Binder is supported only on Windows)
<b>Database Formats</b>	
Access (versions through 2.0)	dBASE (versions through 5.0)
DataEase (Version 4.x)	dBXL (Version 1.3)
Enable (Versions 3.0, 4.0 and 4.5)	First Choice (versions through 3.0)
FoxBase (Version 2.1)	Framework (Version 3.0)
Microsoft Works (Windows, versions through 4.0)	Microsoft Works (DOS, versions through 2.0)
Microsoft Works (Mac, versions through 2.0)	Paradox (DOS, versions through 4.0)
Paradox (Windows, versions through 1.0)	Personal R:BASE (Version 1.0)
R:BASE 5000 (versions through 3.1)	R:BASE System V (Version 1.0)
Reflex (Version 2.0)	Q & A (versions through 2.0)
SmartWare II (Version 1.02)	MIME Text Mail
<b>Other Formats</b>	
Microsoft Works (Mac, versions through 2.0)	Paradox (DOS, versions through 4.0)
Paradox (Windows, versions through 1.0)	Personal R:BASE (Version 1.0)
R:BASE 5000 (versions through 3.1)	R:BASE System V (Version 1.0)
Reflex (Version 2.0)	Q&A (versions through 2.0)



## About Symantec

Symantec is the world leader in providing solutions to help individuals and enterprises assure the security, availability, and integrity of their information.

Headquartered in Cupertino, Calif., Symantec has operations in more than 40 countries.

More information is available at [www.symantec.com](http://www.symantec.com).

For specific country offices and contact numbers, please visit our Web site. For product information in the U.S., call toll-free 800 745 6054.

Symantec Corporation  
World Headquarters  
20330 Stevens Creek Boulevard  
Cupertino, CA 95014 USA  
1 408 517 8000  
1 800 721 3934  
[www.symantec.com](http://www.symantec.com)

Symantec and the Symantec logo are U.S. registered trademarks of Symantec Corporation. Enterprise Vault is a trademark of Symantec Corporation. Microsoft, Outlook, SharePoint, and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. All other brand and product names are trademarks of their respective holder(s). Copyright © 2006 Symantec Corporation. All rights reserved. Printed in the U.S.A.  
03/06 10568566