

A Comparison of Basis 8.3 with DOE/LSN Requirements

ISRI Staff

Technical Report 99-07
Information Science Research Institute
University of Nevada, Las Vegas

December 22, 1999

1 Introduction

The following document reviews Opentext's BASIS System with respect to the Department of Energy's (DOE) search and retrieval requirements for the Licensing Support Network (LSN). We used a "hands-on" evaluation to determine how closely this system meets each requirement.

1.1 Company and Product Information

BASIS, originally a product of Information Dimensions Inc. (IDI), is now sold by Open Text Corporation. Open Text was founded in 1991 and has made several acquisitions including IDI. With this acquisition, the company secures 42.7% of the market share of worldwide enterprise document management installations.

BASIS was commercially available in 1986 so it is considered one of the oldest information retrieval systems. It is referred to as an *extended relational database* because BASIS stores both relational type records as well as full text documents. The BASIS system is best suited for applications such as records management, litigation support, content archives and corporate memory management.

BASIS software requirements include:

- Netscape Enterprise Server (3.61 for Solaris)
- Frame capable browser such as Netscape or Microsoft Explorer

2 Collection Preparation

Installation of BASIS using the supplied install program was very smooth and easy. Unfortunately, the installation of the WEBserver Gateway suffered from multiple bugs in the install script which ISRI had to track and fix.

Loading a collection into BASIS is a fairly arduous procedure requiring the use of several separate tools, each with their own set of commands to learn. For this reason, a trained BASIS expert should be considered a prerequisite for any site planning to use BASIS for its underlying record and document management system. It took ISRI staff approximately 2 man-weeks to learn the required BASIS commands, load the sample collection, and enable the WEBserver Gateway.

Surprisingly, there is no straightforward method for loading HTML documents. Instead, a new collection had to be built using BASIS Generalized Markup Language (BGML) to specify header fields for the record, and using plain text for the body of the document. An example file appears in Figure 1. Once the database was defined and created, these BGML files can be loaded using the High Volume Update (HVV)

utility. Eventually, OpenText did provide a method for loading HTML files. This method involves pre-processing the files with a program that wraps each HTML tag with the BGML HTAG tag so for example “<h2>Abstract</h2>” becomes “<HTAG BREAK=N VAL='h2'>Abstract<HTAG BREAK=N VAL='/h2'>”.

```
<#FIELD NAME=DOCID>0100</#FIELD>
<#FIELD NAME=DOCUMENTTYPE>Publications</#FIELD>
<#FIELD NAME=TITLE>Role of Colloids in Nuclear Waste Disposal</#FIELD>
<#FIELD NAME=AUTHOR>Avogadro, A De Marsily, G</#FIELD>
<#FIELD NAME=ABSTRACT>Aspects of formation and characterization of a
radioactive colloidal fraction released by the waste form or produced
by association with microcolloids naturally existing in ground water
or produced either by corrosion of container material or by
degradation of backfill material are discussed. A filtration model
has been developed in order to describe colloidal transport under
field conditions. Comparison between data obtained with laboratory
column experiments and theoretical evaluations is presented.</#FIELD>
<#FIELD NAME=PAGECOUNT>11</#FIELD>
<#FIELD NAME=Keywords>Colloid Geochemistry Radionuclide Migration
Corrosion Radioactive Waste Canisters Transport Models Backfill
(Repository) Hydrogeochemistry Leaching (Geochemical) Mathematical
Models </#FIELD>
<#FIELD NAME=PUBLICATIONDATE>19840000</#FIELD>
<#FIELD NAME=DOCUMENTSUBTYPE>Conference Papers</#FIELD>
<#FIELD NAME=TEXT>
```

00 /00 c.

495

Ti-IS HOLE OF COLLOTOS Ii\ RL'LL:AR CASTE DISPOSAL
[...]

```
</#FIELD>
```

Figure 1: An example BGML file

If not for bugs in the installation script, the WEBserver Gateway would have been straightforward to use. The WEBserver Gateway is pre-configured with three different interfaces for applying BASIS queries to any BASIS database. Unfortunately, another bug exists which made it necessary to restart the web server anytime there is a period of 5 hours when no query is made. OpenText technical support stated that this bug would be fixed in version 8.4.

3 Requirements Proficiency

3.1 General Requirements

(R) Year 2K Compliance. DOE-LSN is to be Year 2000 compliant.

1. Opentext *does not* state that BASIS or its Techlib products are Year 2000 Compliant. Instead, they've released a *Year 2000 Readiness Disclosure Statement* that points out problems that some versions may have with Y2K. This section lists these issues but does not review them in detail. Please refer to the Opentext website at <http://www.opentext.com/year2000/> for full disclosure.

Their disclosure states:

(a) **BASIS Document Manager**

- i. Date fields (field USAGE=DATE) are stored internally as CCYYMMDD. If user-supplied data or system supplied data is corrected, dates will properly calculate date data.
- ii. BASIS will convert the date to the eight-digit ISO form by applying the "current century" inference rule. For example, on December 31, 1999, an input date of the form 10/22/98 will be stored as 19981022 in BASIS. However, on January 1, 2000, an input date of the form 11/15/98 will be stored as 20981115 in BASIS.
- iii. Any site using a BASIS application that accepts date data in six-digit form (i.e. two-digit years) must take corrective action. The BASIS formats DATE7, DATE8, DATE9, DATE11, DATE18, DATE20, DATE22, DATE25, and DATE26 process the year in two digits and are not recommended for data entry or update.
- iv. BASIS versions prior to 8.2.3 should plan to shutdown and restart all BASIS applications and BASIS Kernels after the century rollover.
- v. In versions prior to 8.2.3, The DATE_REFORMAT function in DMFQM and DMRW does not correctly apply the century inference rule when processing six-digit dates.

(b) **Use of Y2 Date Forms, Record Update Actions, BASIS Find Command**

- i. In some situations, century values are changed, after a REPLACE action.
- ii. In certain usages of BASIS Find, the same FIND command will produce DIFFERENT result sets in 1999 versus 2000.

(c) **BASIS Supplemental Tools**

- i. Since these products are in "Phone Support Only" status, Open Text does not intend to fix customer reported problems, Year 2000 originated or otherwise.

(d) **BASIS Techlib**

- i. Versions of Techlib on NT may have problems with Y2 output format.
- ii. User views of data within the TLP application program use an abbreviated (Y2) date format (MM/DD/YY). BASIS extends user-supplied input data, host operating system services data, and application-calculated data processed through these views by applying the current century (e.g. 10/31/96 is extended and stored as 19961031). Application logic errors are introduced with the current century default when TECHLIBplus calculates future dates.
- iii. Open Text recommends that customers upgrade existing TECHLIBplus applications to BASIS Techlib (version 8.2 or later) before January 1, 2000 to avoid Year 2000 compliance issues.
- iv. The BASIS Y2 date output formats are DATE7, DATE8, DATE9, DATE11, DATE18, DATE20, DATE22, DATE25, and DATE26. Database UPDATE operations using these formats may not function properly over the century change.

(e) **BASIS K and Below**

- i. BASIS product versions through K.9 ("BASIS K") do not properly manage dates and may generate incorrect values or invalid outputs.

Open Text is providing this information to assist you in understanding and addressing your Year 2000 challenge.

(R) **Collection Size.** DOE-LSN must accommodate the anticipated size of 1,000,000+ documents containing 10,000,000+ text pages and images.

REQUIREMENT NOT TESTED

(R) Internet Accessible. DOE-LSN must be accessible on the Internet.

1. BASIS satisfies this requirement.
2. Although the legacy curses-based interface is still prevalent, BASIS does include a web-based interface to the system which can be accessed by clients anywhere on the Internet using ordinary web browsers such as Netscape Communicator and Microsoft Internet Explorer.
3. An API for the system is included, making it possible to create other, non web-based methods of accessing the system across the Internet.
4. There are factors that may affect accessibility of the system, all of which are considerations for any Internet communications, not just for BASIS. Here are a few:
 - Internet firewalls. If communications between the client and server must pass through a firewall device, it is possible that communications may be restricted due to security policies.
 - Web proxy servers. Widely used especially by large ISPs, these devices help to reduce bandwidth demands by caching local copies of frequently accessed web pages. Since BASIS is such a complex system, it is possible that minor problems could arise for users who are forced to go through proxy servers. In fact, in our test environment, problems were noticed when attempting to run queries using the web-based interface when going through a Squid proxy server. The problems went away after changing Netscape Communicator's proxy setting to "Direct connection to the internet."
 - Network bandwidth. The transfer of document images and other types of data will make considerable demands on available network bandwidth. Adequate bandwidth on the server side must be available for BASIS to serve requests from all users who wish to access the system.
 - Network latency. Latency is the apparent delay between the time some data is transmitted from one point on the network to when it is received at another point. With an interactive system such as BASIS, it is important to consider the effect of network latency on the usability of the system. For example, if the average delay is too great between when a user clicks something and a response is received, many users may feel that the system is too frustrating to use.

5. Overall Impression

Although BASIS has adequate support for connectivity across the Internet, in working with the system, occasionally some relics of the system's pre-Internet versions are exposed. They tend to give the system a cumbersome feel.

(R) Windows/Windows NT. DOE-LSN must be usable by clients on Windows and Windows NT operating systems.

1. BASIS satisfies this requirement.
2. The BASIS system is usable by Windows-based clients in a number of ways, including through a web browser, through a programmatic interface such as one that is available for Visual Basic, and through the BASIS Desktop application which is implemented in Visual Basic. Only the first method was considered in this evaluation.
3. The most widely usable interface to the BASIS system for Windows systems would be to implement a web-based interface. This would allow virtually any Windows system that can run a web browser to access the BASIS system.
4. **Overall Impression**

There are several methods for accessing the BASIS system available to users of Windows systems. A web-based interface, implemented using the BASIS WEBserver Gateway, could provide an interface sufficient for users of such systems.

(B) **Platforms/Operating System.** DOE-LSN should run on one of the following platforms: Windows NT, Sun Solaris, Alpha Unix.

1. BASIS satisfies this requirement.
2. BASIS 8.3 is available for the following platforms:

Operating System	Platform
Digital OpenVMS	DEC Alpha
Hewlett Packard HP-UX 11	HP-9000
IBM AIX 4.3	RS/6000
Microsoft Windows NT w/ SP4	Intel x86
Sun Solaris 2.6	SPARC

Notes:

- Not all functionality is available in the Windows NT product. Specific capabilities should be verified before selection of a server platform.
 - Only Sun Solaris was tested in this evaluation.
3. There are no known issues which would affect performance on this platform.

(R) **Concurrent Users.** DOE-LSN shall support up to 150 concurrent users. [LSS2-064]

REQUIREMENT NOT TESTED

3.2 Querying Requirements [LSS2-011]

(R) **Query for Document.** The DOE-LSN shall provide the capability to query the system for a list of all documents that meet the query criteria and sort the displayed list on the basis of selected displayed fields or relevancy to the query. [LSS2-011]

1. BASIS returns a list of documents that meet the query criteria.
2. *Query for Document* is an integral part of the BASIS system. The BASIS system provides two default user interfaces:
 - (a) The *FQM Query form* is shown in Figure 2. This form consists of a simple text box allowing the user to enter his query using the BASIS query language - Fundamental Query and Manipulation (FQM) language.
 - (b) The *Field Template form* is shown in Figure 3. In this form, the user can enter his query using terms and phrases for text fields or enter numeric and date values for these field types. The search criteria can be combined by selecting AND or OR but query combinations are limited by allowing the selection of only one field connector for the entire form.
3. Ascending and Descending sort order can be specified for a selected field in BASIS. BASIS also provides the capability to sort by relevancy rank.
4. Querying the system and sort order is well documented in the BASIS documentation.
5. **Overall Impression**
BASIS has the capability to return a list of documents that satisfy a given query. Furthermore, BASIS is able to sort the returned documents by either rank or ascending/descending order on the field of choice.

(R) **Query Header.** The DOE-LSN shall provide the capability to query the system by specifying the content of one or more header fields to obtain a list of all documents that satisfy the query. [LSS2-011-1]

Comment: The search will be able to sort appropriate fields, such as date, accession number, etc. in ascending or descending order. It is anticipated that the DOE-LSN will allow the user to select and search multiple bibliographic header fields.

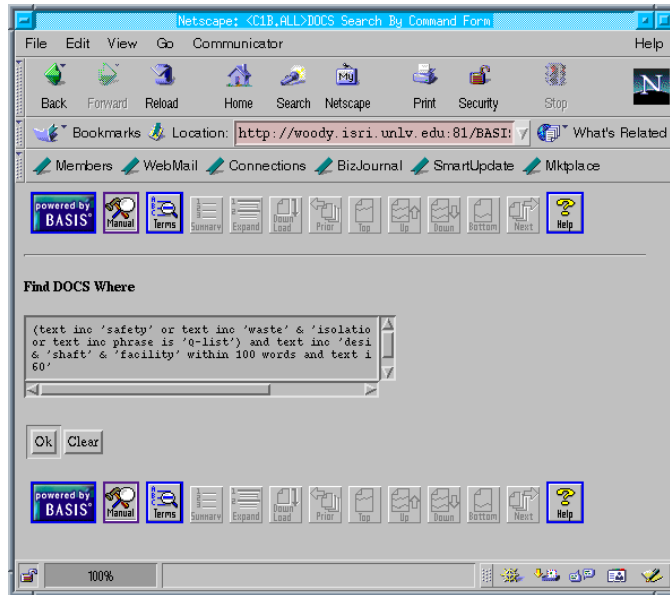


Figure 2: FQM Query Interface in BASIS

1. BASIS provides the ability to search fields and return documents relevant to the fielded search. Any number of header record fields may be selected in a single search.
2. Field searching can be accomplished in either of the available interfaces. Using FQM, the field names must be known to the user. An example header record query appears in Figure 4.
3. **Overall Impression**

Although the default field template form is restrictive in its ability to combine fields, the interface can be modified to allow a more flexible means of querying header fields.

(R) Query Text. The DOE-LSN shall provide the capability to query the system by specifying one or more character strings in the full text of the document to obtain a list of all documents that satisfy the query. [LSS2-011-2]

Comment: Describe any query optimization techniques used by your system.

1. BASIS provides the capability to specify character strings to search the full text of the document collection. Text searching is an integral part of the BASIS system.
2. BASIS applies Boolean-like, exact match searching using FQM. FQM includes the standard Boolean operators (AND, OR, NOT), as well as other operators for phrase searching and proximity searching.
3. In BASIS, query optimization is done through several environment variables that affect query processing of the FIND command. These optimization variables fine-tune the command processing based on specific needs. Depending on settings, the FIND command processor will evaluate differently the restrictions in the AND command. For example (from BASIS documentation):

DM_FIND_AND_THRESHOLD_MAXNR if this variable has a value of n , an AND operation that would result in more than n records will force the FIND command processor to not consider evaluating any restrictions in the AND operation by scanning the records. All indexes will be used in the evaluation.

If this threshold variable has a value of n , an AND operation that may result in fewer than n records but more than **DM_FIND_AND_THRESHOLD_NR** records will force

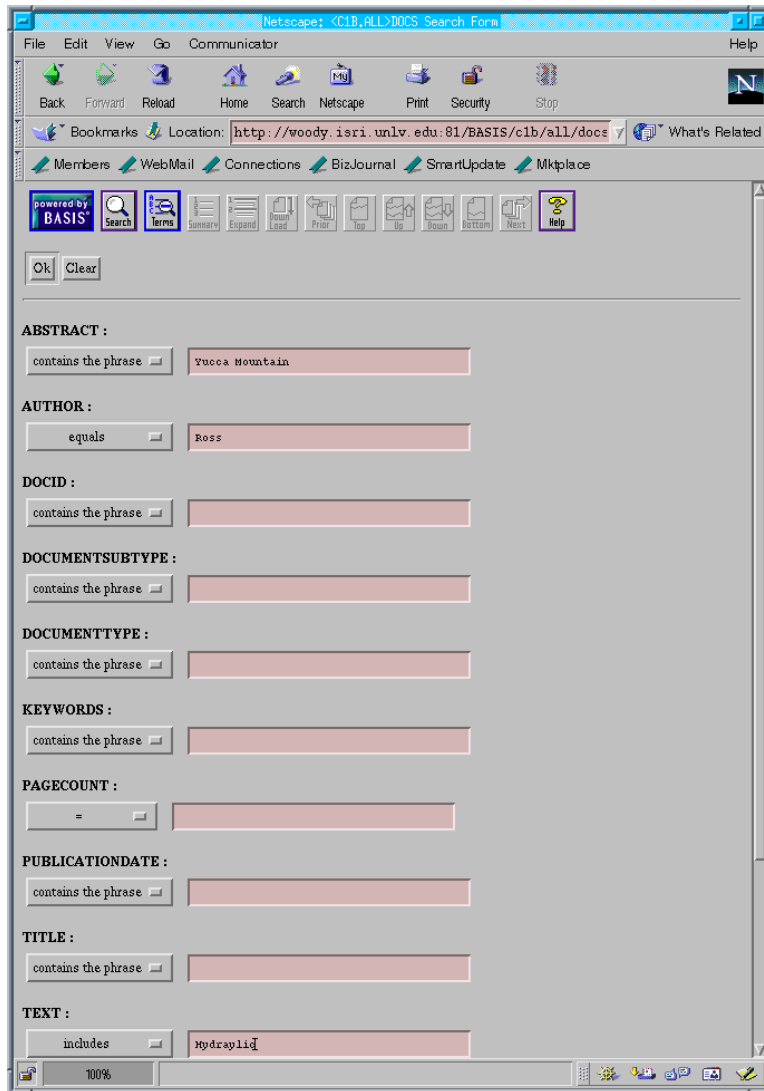


Figure 3: Field Template Form in BASIS

```
FIND documents WHERE text INC PHRASE LIKE 'Sandia National Laboratories' + & 'Ross' &
'waste repository' & 'Yucca Mountain' OR + text INCLUDE PHRASE LIKE 'SNL' & 'Ross' &
'waste repository' & + 'Yucca Mountain' ORDER BY docid
```

Figure 4: FQM Header Field Search

the FIND command processor to consider evaluating any restrictions in the AND operation by scanning the records.

DM_FIND_AND_THRESHOLD_NR specifies the maximum size of the final result set. The FIND command processor will always use this to turn off some of the indexes so it may scan the records in order to evaluate part of an AND operation.

In addition, there are several environment variables associated with the cost of accessing the index. Examples include:

DM_FIND_CF_PER_INDEX specifies the cost factor for accessing one reference in the index. This is used together with **DM_FIND_CF_PER_RECORD** in deciding whether it is better to access indexes or to access records to satisfy the FIND command.

DM_FIND_CF_PER_RECORD specifies the cost factor for accessing one record in the database and scanning data to check retrieval tests. This is used in deciding whether it is better to access indexes or to access records to satisfy the FIND command.

4. Since the text stream of a document is clearly distinct from its header fields, a user must explicitly request that each field (including the text field) be searched.
5. Searching full text is well documented in BASIS' documentation.

6. Overall Impression

BASIS meets this requirement. Header fields are not automatically searched when performing searches on the text stream of a document.

- (B) **Text Query Parameters.** The DOE-LSN shall provide the capability to specify single and multiple character wildcards, to utilize proximity searching, and root searching as part of a full-text query and to combine multiple result sets. [LSS2-011-3]

Comment: The intent of this requirement is to allow users to combine any of the previously executed searches that were performed during the same search and retrieval connection.

1. BASIS provides multiple character wildcarding and other parameters as shown in Table 1 to aid searching.
2. Proximity searching is available in BASIS. Without text markup, BASIS provides both word and sentence proximity searching. Other *context units* can be defined by the Database Administrator.
3. BASIS does not provide root searching in the complete sense; BASIS only performs a simple plural to singular algorithm (s removal) if the SINGULAR parameter is set.
4. BASIS also provides the ability to:
 - control word break characters.
 - control sub break characters.
 - control case searching.
 - convert control characters to blanks.
5. BASIS does allow the administrator to apply a *filter* to indexed terms. This would allow some level of processing to be performed (like stemming) before the word is stored in the inverted file. This same filter would be applied to query terms at search time. Only the unfiltered version of the field, not the filtered version, can be displayed.

6. Overall Impression

BASIS provides a minimal set of text search parameters for searching the full text of the documents in the LSN. Any advanced text search algorithms would have to be added using filters.

- (R) **Query Header and Text.** The DOE-LSN shall provide the capability to query the system by specifying a combination of header field values and the text query parameters from the full text of the document to obtain a list of all documents that satisfy the query. [LSS2-011-4]

Parameter	Example	Description
*	'JACK'*	matches from 0 to 15,000 characters
#	'museum of # art	word length wildcard

Table 1: Wildcard Searching in BASIS

Operators	Search Values	Description
>, >=, <, <=	numbers, character strings, dates, fields, time service and security functions	Compares values as indicated.
=, ^ =	numbers, character strings, dates, fields, time service and security functions, ranges, character patterns, list of numbers, NULL	Compares values as indicated.
INC	Character strings, character patterns, list of character strings, character patterns	Operator used for textual fields
ALL WORDS	phrases, list of phrases	Finds documents where search terms appear in the same context unit in any order.
ANY WORDS	phrases, list of phrases	Finds documents with any of the key words in any order.
PHRASE IS	phrases, list of phrases	Finds documents with exact phrase.
PHRASE LIKE	phrases, list of phrases, phrase patterns, list of phrase patterns	Finds documents with similar phrase allowing some variation.

Table 2: Query Operators in BASIS

1. BASIS satisfies this requirement.
2. FQM, the BASIS command language, accommodates header field and text searching within a single query. This is also supported in the Field Template interface provided with the system as shown in Figure 3.
3. The FIND command allows the operators listed in Table 2 for matching search requests. As noted in the table, there are restrictions on the type of search values to which these operators can apply.
4. BASIS documentation gives information on the syntax and use of these search methods.
5. **Overall Impression**

BASIS provides the capability to search header and text fields simultaneously and has operators for retrieving various forms of textual phrases. Some operators though have inconsistencies making these operators difficult to learn.

- (R) Provide Query Status.** The DOE-LSN shall provide the user an indication of the query status during a query and allow the user to terminate queries in process without terminating the session or losing previous result sets. [LSS2-011-5]

Comment: It is always possible to construct a query so broad that it results in an unmanageable results list. Users should be able to determine that an ongoing query is too-broad and terminate the query in process. An indication that the session is still connected and that the query is working is adequate.

1. BASIS web interface does not provide any specific query status indicator while the system is searching other than what is displayed by the browser. The browser indicator is specific to the browser (i.e. Active "N" icon in Netscape Navigator, and spinning "e" icon in Internet Explorer). While the request is in progress, the status line at the bottom of the browser gives information about page download in bytes retrieved.

2. The download of any page transfer can be terminated by selecting the **Stop** icon on the browser, but again, this is a browser feature and may differ from browser to browser. This does not interrupt the current session and does not affect any *previously saved* result sets.
3. Since these are browser features rather than an integral part of BASIS, no documentation is provided.
4. The interface does not indicate the total number of documents retrieved after a search is complete. However, the user has the ability to display the last page of the results list showing the last document retrieved. The interface could be modified to display this number if required.
5. The BASIS environment variables allow settings for the time limit of query processing. `DM_FIND_TIMEOUT` variable specifies the maximum CPU time (in seconds) allocated to a single `FIND` command. Any `FIND` command that exceeds this usage will be aborted. This variable is set by the system administrator. In addition, the user can dynamically indicate a timeout value for a query by setting the `TIMEOUT` parameter on the `FIND` command.

6. Overall Impression

Most users will be familiar with their browser and will know how to observe and use these features. The user can easily determine whether the query is still in progress or stalled, but there is no other query status information.

- (B) Query Assistance.** The DOE-LSN shall provide interactive capabilities to assist the user in retrieving documents when the field values that uniquely define the documents are not known to the user. [LSS2-011-6]

Comment: Examples might include synonym processing, thesaurus, natural language queries, or other search aids. Because a variety of approaches are used in the commercial market, no one approach is specified.

Natural Language BASIS is a relational database system for short numeric and character fields as well as long textual streams. *FQM* is its command-oriented language used to search both relational as well as text fields in BASIS. Querying in *FQM* is similar to SQL querying but provides its own set of commands and parameters. BASIS does provide the ability to apply a natural language interface (as shown in Section 3.3). This interface should be designed carefully so as not to lose the functionality of the *FQM* language.

Thesaurus: BASIS does provide the Houghton Mifflin Roget's Electronic Thesaurus. The Roget's Thesaurus differs from other BASIS thesauri in that it is already built. It is not supplied with the system, but can be obtained at an additional fee.

BASIS also *accommodates* the building of collection specific thesauri. BASIS dedicates a separate module (TM) to thesaurus construction, maintenance, and storage. The thesaurus structure is hierarchical and requires hundreds of man hours for its construction and upkeep. This is the *only* thesaurus format that can be used in association with a BASIS database. Once a BASIS thesaurus has been associated with a document collection, the user is also given the ability to browse the thesaurus. This feature can be helpful to the user especially when the collection is domain specific.

Soundex: BASIS provides Soundex encoding which could be used to help users locate misspellings.

Manipulating Result Sets: BASIS allows the user to explicitly combine query result sets using boolean operators. BASIS also gives the user the capability to restrict his search to previous result sets. This gives the user the ability to refine and narrow his search.

Feedback: BASIS does provide a form of feedback using the LIKE function. A complete result set, selected members of a result set, or context within members of a result set can all be used as feedback to find similar documents.

Number and Date Searching: BASIS allows number and date searching but only on relational fields that are designated as numeric and date fields, not in text streams.

Overall Impression

BASIS provides standard query assistance tools. The most notable tool is BASIS' ability to manipulate query result sets. The system's use of feedback could also prove useful as long as this ability is an easy to use selection built into the user interface.

3.3 Display Capabilities [LSS2-012]

1. **NOTE:** All display capabilities are dependent on the browser and interface to the system. The default interface provided by BASIS was used in this evaluation.

(R) **Display Document** The DOE-LSN shall provide the capability to display a document. [LSS2-012]

1. BASIS provides the ability for a user to select a document to be displayed from the results list of a query.
2. A document can be viewed after a query has been run or by selecting an occurrence of a search term from the index. The index displays a list of the term's occurrences (See Figure 5). A selection of one of these terms displays the set of documents in which the searched term occurs.
3. The document has been broken into "pages" which can be accessed in order. These pages are *not* equivalent to the hard copy pages.
4. **Overall Impression:**
BASIS satisfies the document display requirement. The document page segmentation does not necessarily correspond to the actual pages of the document.

(R) **Display Header** The DOE-LSN shall provide the capability to display the header of a document. [LSS2-012-1]

1. BASIS provides several ways for displaying a document header.
2. When a search is run, BASIS displays the rank and the title field of the retrieved documents in the results list. These fields though can be any or all of the fields of the header record.
3. Document headers are also displayed as part of the document.
4. The default interface provided with BASIS includes an **expand** option to view all header records for the documents in the results list. See Figure 6 for displaying header records in BASIS.
5. **Overall Impression**
BASIS provides for header field display. This display can be modified by the system programmers to include the header fields that would most benefit the user.

(R) **Display Text** The DOE-LSN shall provide the capability to display a page of text of a document. [LSS2-012-2] Text Format: The text representation of material in DOE-LSN shall be page delimited ASCII text. [LSS2-056]

1. BASIS provides this capability.
2. When a document is selected from the results list, pages of the document are displayed in HTML text representation. These pages do *not* necessarily correspond to the hard copy pages of the document.
3. **Overall Impression**
BASIS satisfies this requirement without modification.

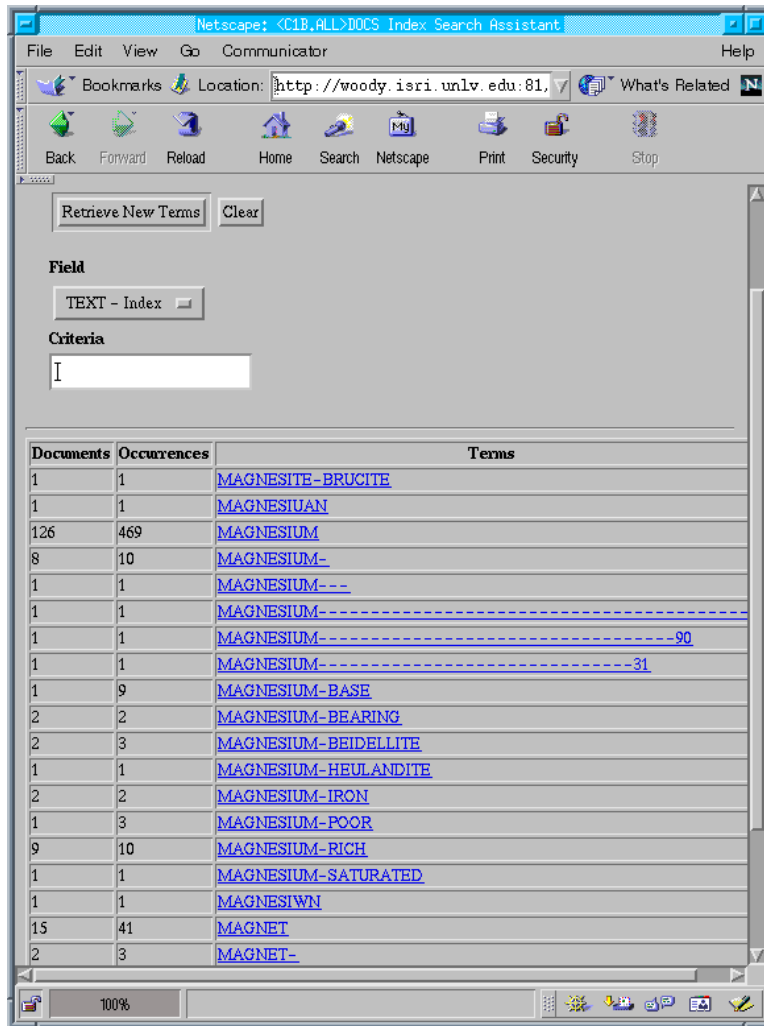


Figure 5: Term Occurrence Allows Document Selection and Display

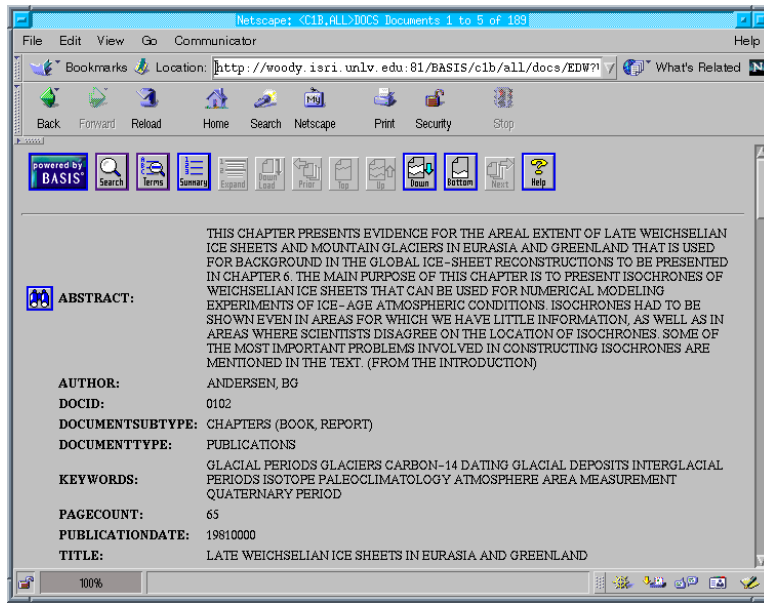


Figure 6: Header Records Displayed in BASIS

(R) Locate Search Terms in Document The DOE-LSN shall provide the capability to locate the terms in the document text that satisfy a full-text query and to move from one term to the next or previous term without displaying intermediate text. [LSS2-012-3]

Comment: This function is performed as the user is viewing the document. It is typically implemented by highlighting the search terms in the document and providing a “go to next term” function that places a cursor at the line or word of the search term.

1. BASIS provides this capability when a document is selected from the results list.
2. Search terms appear in bold surrounded by icons allowing for jumps between consecutive terms. This allows for immediate location of search terms in the document. These search term icons are shown in Figure 7.
3. **Overall Impression**
BASIS provides term highlighting and the ability to move forward and backward between consecutive hits to aid in locating search terms.

(R) Display Image The DOE-LSN shall provide the capability to display the images of a document, page by page, including full page views of the images of 8-1/2 by 11 inch pages up to E-size pages.[LSS2-012-4] Image Formats: The electronic image of documentary material in DOE-LSN shall use Aldus Tagged image File Format (TIFF) Group 4 for bitonal images and Joint Photographic Experts Group (JPEG) for color and gray scale images. These formats are part of the Adobe TIFF I Version 6.0 representation. Adobe TIFF is an industry standard developed and put into the public domain by Adobe. [LSS2-057]

1. Using the default interface, BASIS did not satisfy this requirement.
2. BASIS uses *converters* to load collections with markup. There is no HTML converter, and to write one would have been an arduous task. For this reason, we were unable to load the HTML collection and in turn, unable to display TIFF images using plugins.
3. BASIS tech support claims image data can be loaded into a stream field that would correspond to a document page. Our document collection though would have to be modified to accommodate this change.

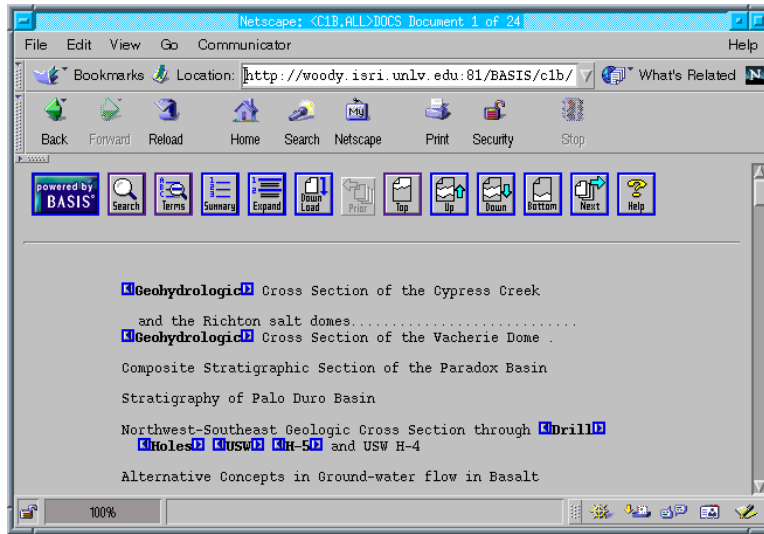


Figure 7: Highlighting Search Terms in BASIS

4. None of the images in our test collection were JPEG images but JPEG viewing is accommodated by most Internet browsers in use.

5. **Overall Impression**

If the ability to load images were available, BASIS would still require a TIFF plugin viewer to satisfy this requirement. Appropriate plugin selection is an important consideration for the LSN.

(R) Image Viewing The DOE-LSN shall provide image viewing for image enlargement, reduction, scrolling, and. [LSS2-012-5]

1. Only with a separate TIFF viewer can BASIS satisfy this requirement. The plugin should be capable of enlarging, reducing, scrolling and rotating the image being displayed.
2. For this evaluation, ISRI installed *AlternaTiff* a TIFF viewer plugin to meet this requirement. This plugin is equipped with buttons which, when clicked, perform the image augmentations listed.

3. **Overall Impression**

BASIS can meet this requirement with the appropriate plugin installed. Appropriate plugin selection is an important consideration for the LSN.

(R) Display Image and Text The DOE-LSN shall provide the capability to concurrently display an image page of a document and its text. [LSS2-012-6]

Comment: There must be a one-to-one correspondence between each page of text and its corresponding page image. This assumes each page will be tagged in the text version.

1. BASIS can satisfy this requirement if 1) A plugin is installed which launches a new window for TIFF image display, or 2) BASIS's user interface is designed to split the current browser into frames to display the image in one frame and the document text in the other.
2. With ISRI's HTML implementation of the collection, TIFF image thumbnails have been embedded at the beginning of each page of text for toggling between the page and its corresponding page image. This implementation demonstrates the ability to obtain one-to-one correspondence between the text page and image page.

3. **Overall Impression** Some solutions to this requirement may be dependent on collection preparation.

(R) **Viewing Options** The DOE-LSN shall allow the user to view the following combinations: 1) header, 2) image, 3) text, 4) header and text, 5) header and image, and 6) text and image.[LSS2-012-7]

1. BASIS interfaces satisfy most but not all display combinations listed in this requirement.
2. To view the header fields of a document, the user must select the expand button at the results list or the listing of a document. One or more specified fields can be listed with the results list.
3. The TIFF image for each document page can be included as a field in the collection table. The user can click on a link to bring up the image of the current text page being viewed. The images have to be incorporated into the documents during collection preparation. This is not an inherent capability of BASIS.
4. The text of the document can be displayed only after a query is run.
5. The header and text combination is displayed automatically when a document is selected from the results list.
6. None of the BASIS interfaces display the combinations: images and header, images and text in a single browser window. Modifications might be possible to accommodate such requirements.
7. **Overall Impression** BASIS satisfies a number of the specified requirements, however, to accommodate all of the requirements, interface modifications are necessary.

3.4 Printing Requirements [LSS2-013]

NOTE:

The printing capabilities of BASIS depend entirely on the browser in use. Printing capabilities were tested extensively using several versions of Microsoft Internet Explorer and Netscape but exceptions may occur.

(R) **Print Document.** The DOE-LSN shall provide the capability to print a document at a local printer. [LSS2-013]

Comment: It is assumed that the local printer is capable of printing the requested document type.

1. BASIS satisfies this requirement through the use of a web browser.
2. To obtain the document for printing in its entirety, the user selects a document from the results list and then selects `download`. A user may also print each web page displayed. This page is *not* equivalent to a hard copy page of the document.
3. Exhibit A shows a printed ASCII document from the BASIS system after the `download` option was selected. It includes the full document. Other document components would have to be printed separately.
4. **Overall Impression**
BASIS depends on the user's browser capabilities to print a document from the DOE-LSN collection. The appearance of the document printout depends on the preparation of the collection (See Section 2 for more details), BASIS' processing of the document, and the selected print option.

(R) **Print Header.** The DOE-LSN shall provide the capability to print a document header at a local printer. [LSS2-013-1]

1. BASIS satisfies this requirement in that it can print a "page full" of headers but not a single selected document header.

2. The `expand` option retrieves all header records for the documents in the results list. Header fields are displayed five to a page. By scrolling through the header records, the user can select which document's header record should be printed.
3. The `expand` option is also available once a selected document is displayed. this selection retrieves the current document's header record first in the list.
4. **Overall Impression**
BASIS provides a nicely formatted way to view and print document headers. With the default interface though a single document header cannot be printed without using cut and paste methods. This could be changed in a customized interface.

(R) Print Text. The DOE-LSN shall provide a user selectable capability to print from one page to all of the text of a document, and any selected ranges of pages, at a local printer. [LSS2-013-2]

Comment: The system must be able to discern pages within a document for printing.

1. BASIS allows printing of the entire document or a selected page.
2. To print the document in its entirety, the user selects the document from the results list and then selects `download`.
3. To print separate pages, the user must select a document from the results list, scroll through the document using the `up`, `down`, `top`, and `bottom` options to reach a desired page. These pages do *not* correspond to the hard copy pages of the document.
4. The default interface does not give the option for printing a range of pages. This could be added to a customized interface.
5. **Overall Impression**
BASIS is able to print a document in its entirety or a single page of a document. Printing ranges of pages of text could be implemented but should be considered carefully to ensure its usefulness in the browser/HTML environment.

(B) Report Generation. The DOE-LSN should provide report generation capabilities for several of the above listed tasks.

1. BASIS provides a report generation facility: the BASIS Report Writer module (RW).
2. RW allows the user to generate reports from the data in the collection table. If data regarding specific requirements is needed, the collection table may have to be modified with additional fields to reflect data usage and updates.
3. Extensive documentation for using this module is available online.
4. **Overall Impression**
BASIS has well-documented report generation capabilities.

(R) Print Standard Image. The DOE-LSN shall provide a user selectable capability to print from one to all images, and any selected ranges of images, of 8-1/2 by 11-inch (or smaller) pages of a document, at a local printer, reduced to a single 8-1/2 by 11-inch paper. This includes the capability of printing an oversized page image, up to E-sized, on a single 8-1/2 by 11-inch sheet of paper. [LSS2-013-3]

1. Since considerable modification of our test collection would have been required to implement this feature , this requirement was not tested for BASIS.
2. Although not implemented, BASIS tech support claims: "images can be loaded as fields in the collection database and accessed through the WEBserver Gateway."
3. If indeed this is possible, then TIFF images would be printable using a plugin like the TMSSequoia TIFF viewer. The user would be required to download a viewer plugin before viewing or printing a standard image. This is not an inherent capability of BASIS.
4. Most viewer plugins have the ability to print the current image to a local printer. Viewers *usually* include the ability to reduce, enlarge, and rotate an image page, so these capabilities should be met in general.

5. The ability to print a *selected range* or *all* page images of a document is usually not a feature of viewer plug-ins.
 6. **Overall Impression**
In addition to modifying the collection database, a plugin viewer may need to be adapted. Selection and distribution of an appropriate plugin together with thoughtful integration of image printing into the user interface would be required to meet this requirement.
- (R) **Print Oversized Image.** The DOE-LSN shall provide the capability to print an oversized page image, up to E-sized, on a single sheet of paper at 100 percent of the size of the original image. [LSS2-013-4]
1. As indicated above, image loading, display, and printing was not evaluated for BASIS.
 2. In addition, printing an oversized image depends on the capabilities of the printer being used.
 3. **Overall Impression**
Since all printing in BASIS would be browser and plugin dependent, printing oversized images would also depend on the selection of an appropriate plugin. Selection and recommendation of the plugin to LSN users is an important consideration.
- (R) **Print Results List.** DOE-LSN shall provide the capability to print some or all of the summary lines of a results list. [LSS2-013-5]
1. BASIS satisfies the ability to print the results list (See Figure 8). The results list may span web “pages.” In this case, each page must be printed separately.
 2. The interface used in our evaluation displays the document rank and the document title, although the abstract could just as easily have been displayed as part of the results list, and therefore printed.
 3. The default interface does allow the results list to be “expanded,” displaying all header record information for all the documents returned in the results list. These pages can also be printed. A single web page contains 5 header records for printing (see Figure 6). To print a specific document header field, the interface would have to be modified.
 4. Allowing the user to select the number of summary lines for print is not an integral part of BASIS.
 5. **Overall Impression**
The design of the interface dictates what is displayed in the results list, and therefore what is printed. Since the interface is customizable, any desired information can be included by modifying the interface.
- (R) **Print Screen.** DOE-LSN shall provide the capability of printing the screen display. [LSS2-013-6]
1. Using the print capabilities of the browser, BASIS can print the “HTML page” displayed. This may or may not define the “screen display.” If the “screen” is defined as just the viewable area within the browser, printing just this portion is not possible unless the full HTML page is in view.
 2. **Overall Impression**
BASIS’ capability of printing the screen depends on the HTML page being displayed.
- (R) **Request Paper Copy.** DOE-LSN shall provide the capability to submit an electronic request for a paper copy of the header, images, or text of a document or of an entire results set, including oversized and color images. [LSS2-014]
1. This requirement’s evaluation is based on the ease with which it could be added to the system since typically it would not be a standard feature of any search system.
 2. Since BASIS is customizable, “Request Paper Copy” could be integrated into BASIS’s interface based on the browser capabilities.

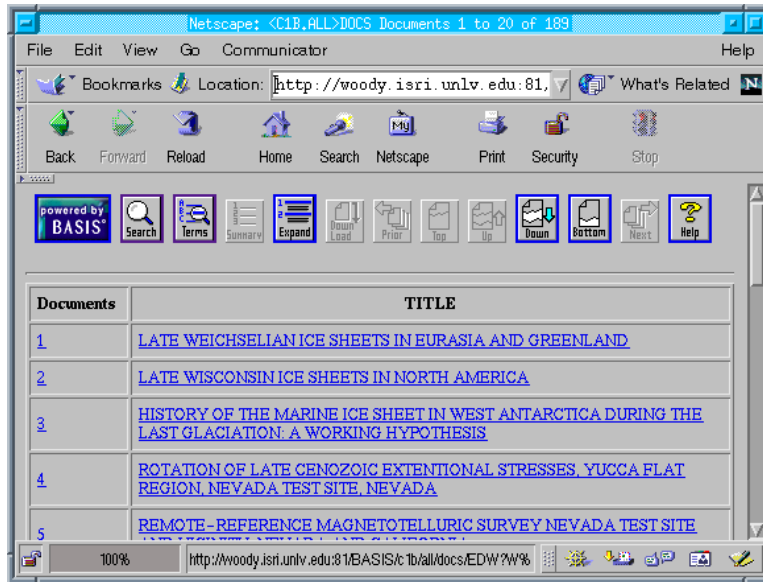


Figure 8: BASIS Results List

3. Overall Impression

BASIS can be customized to accommodate this requirement.

- (R) Process Paper Copy Requests.** DOE-LSN shall provide the capability to receive and read an electronic request for a paper copy of a document and print the requested copy.

Comment: This is not anticipated to be a highly automated function. The requested body must be able to receive requests and print out the requested document. The rest of this function may be procedurally implemented.

1. This requirement is the receiving end of the “Request Paper Copy” requirement. Again, it would not be a standard feature in most search systems. In evaluating this requirement, we have assessed its ease of implementation.
2. This feature can be integrated into the user interface based on browser capabilities.

3. Overall Impression

While this is not a standard feature, it can be implemented in BASIS.

3.5 System Administration Requirements

- (R) Monitor System Status.** DOE-LSN shall provide authorized users the capability to monitor the status of the system and communication components and to interrupt, restrict, or disable capabilities in order to optimize use of system resources. [LSS2-033]

1. This requirement is satisfied by the BASIS product.
2. Fine-grained control over specific non-SuperUser activities for certain users is provided in BASIS by the use of the Data Management System Administration (DMSA) utility in conjunction with the Authority Database (ADB). The ADB controls to what extent a user has access to a library.
3. BASIS gives the system administrator the ability to create user accounts with different privileges. Each user account can have a combination of the following sign-on privileges: SA (can control the ADB, in addition to the rest of the capabilities), OP (can control execution of the Data Management “kernels,” REG (can manage user accounts),

CREATE (can create databases using the DMDBA utility), and WIZ (allows system programmers access to special functions).

4. BASIS allows users with the appropriate access to a library to schedule file conversion and indexing processes. During conversion and indexing, the administrator or file owner can control whether the library will remain available for access by either providing a “parallel copy” of the library or restricting access to the library.
5. BASIS allows users with ‘SA’ or ‘OP’ access to place limits on the amount of memory and other resources that will be used for indexing and searching on a per kernel basis. Each kernel can manage one or more libraries. Depending on the kind of limit changes, a restart of the kernel process may be required. Changes to parameters associated with indexing of the collection may either require a re-indexing of the collection or a dump/reload/rebuild process.
6. BASIS provides extensive logging and user audit trails for most operations. Users with “SuperUser” access can enable or disable certain logs, and in some cases change the amount of information that is logged. Changes to configuration requires a restart of the kernel process.

7. Overall Impression

BASIS provides extensive monitoring capabilities in the form of log files and audit trails. Users may be granted or denied access to certain libraries. The amount of system resources consumed by a collection of libraries can be controlled by configuration of the controlling kernel process.

(R) Monitor Session Activity. DOE-LSN shall provide the capability for an authorized user to monitor user session activity levels and identify and cancel queries or other system activities. [LSS2-033-1]

1. This requirement is satisfied by BASIS.
2. BASIS allows users with ‘SA’ or ‘OP’ access to enable audit logs which record general system activity for all users. Detailed reporting of submitted queries and results may require customization of the interface.
3. BASIS allows users with ‘SA’ or ‘OP’ access to cancel document conversion or indexing processes.
4. BASIS supports interruption of active queries by the user. Users with the ability to control a kernel may disconnect a user or program connected to a particular kernel or library.

5. Overall Impression

BASIS has extensive monitoring and control capabilities.

(R) Database Administration Tools. DOE-LSN shall provide authorized users the capability to assess the availability, integrity, and performance of the databases of the the DOE-LSN, including those pertaining to the storage of document header fields, text, and image data, and adjust database performance parameters or restrict or disable database features in order to optimize system performance.

1. BASIS satisfies this requirement.
2. The BASIS administration tools provide extensive database capabilities. The online documentation describes parameters that control the operation of BASIS and its databases. Any user with administrative or DBA privileges to the system or databases may use the following utilities to review these parameters and adjust them directly:

DMSA This module allows the System Administrator (SA) to control access to BASIS and to specify who is allowed to create databases. DMSA maintains an authorization database for this purpose. DMSA also allows you to check information resident in the Kernel when problems occur and to start and stop the kernel. DMSA allows the SA to monitor and adjust resource parameters (i.e. memory and disk usage) on a per kernel basis.

DMCCF The Communications Control Facility, used by the UNIX and VMS operating systems, tunes the Inter-Process Communications (IPC) software to the needs of your site.

3. Overall Impression

BASIS provides several system and database administration tools to monitor and adjust system and database performance.

3.6 Internet Requirements

(R) Web Server Interface. DOE-LSN must interface with a Web Server for querying the system and returning query results.

1. BASIS satisfies this requirement.
2. BASIS has a template and CGI-based package called the BASIS WEBserver Gateway that provides a web-based interface to a BASIS database with little or no programming. The package is integrated with the Netscape Enterprise web server and the Microsoft web server. The documentation also mentions that Apache and Apache-like web servers can also be used.
3. It is also possible to design and build alternative web-based interfaces to the BASIS system using the BASIS OpenAPI. Requirements may dictate that more customization is necessary than the BASIS WEBserver Gateway can provide. Using the API, a completely different interface can be built that interfaces with a web server, using the CGI mechanism.

4. Overall Impression

The BASIS WEBserver Gateway and OpenAPI make it possible to implement web-based interfaces. All the tools are present that are necessary for constructing elaborate systems for querying a database and returning query results.

(B) CGI and Perl5. DOE-LSN should use the CGI Standard and be accessible from the Perl5 programming language.

1. BASIS satisfies this requirement.
2. The BASIS WEBserver Gateway uses the CGI standard to provide a web-based interface to the BASIS system. Requests from client browsers are received by the web server, handed off to the WEBserver Gateway CGI program, queries are made to BASIS through the OpenAPI interface, then the results are substituted for tags in template files before the final HTML text is served back to the client. The BASIS WEBserver Gateway literature further states that:

Popular Web development interfaces and tools such as Java, JavaScript, ActiveX, Active Server Pages, CGI, or browser plug-ins can be used to enhance interface features.

3. As an alternative to the BASIS WEBserver Gateway, CGI standard conforming applications can also be built from scratch using the BASIS OpenAPI as an interface to the BASIS system. Exactly how such an application implements an interface to the BASIS system, and whether or not it strictly adheres to the CGI standard is entirely up to the developer.
4. Perl5 support is provided by the BASIS Perl DBI Driver, a supported product. The driver binaries and source code are available on an open source basis. The driver enables access to the BASIS system from the Perl language, using the familiar DBI interface. The BASIS Perl DBI Driver v1.2 documentation states that:

...the BASIS Perl DBI Driver provides access to a heavy majority of both the DBI spec and the BASIS OpenAPI capabilities and is expected to meet the needs of most BASIS application developers.

5. It must be noted that the BASIS Perl DBI Driver is implemented using the Perl DBD::JDBC module. As a result, the BASIS JDBC Driver must be purchased, installed, and configured before the BASIS Perl DBI Driver can be used.
6. As an alternative to the above scenario, it should be possible to develop a native Perl BASIS DBD module using the BASIS OpenAPI. Once such a module exists, it would no longer be necessary to have the BASIS JDBC Driver.
7. Another alternative for gaining access to BASIS from Perl would be to write glue code to provide access to the BASIS OpenAPI routines directly from the Perl language. This would be a significantly less desirable situation than using the DBI/DBD modules, but might be necessary if the DBI/DBD modules are too restrictive for a certain application.

8. **Overall Impression**

The BASIS WEBserver Gateway product provides an interface that conforms to the CGI standard, and alternative web-based CGI-compliant interfaces can be built using the BASIS OpenAPI. Perl5 support is not included with the system but is available as an add-on feature.

(B) ODBC/JDBC Compatibility. DOE-LSN should be accessible using Open Database Connectivity (ODBC) and Java Database Connectivity (JDBC).

1. BASIS satisfies this requirement.
2. BASIS provides ODBC access via the BASIS ODBC Driver, implemented using the BASIS OpenAPI. It is not clear for what platforms it is available and it is also not clear what the cost of the driver is.
3. BASIS recently announced the availability of the BASIS JDBC driver. The driver is priced at \$15,000 per BASIS server. Again, it is not clear for what platforms the driver is available. Open Text Corp. states in their October 5, 1999 press release that:

The BASIS JDBC Driver enables Web developers to build applications that take advantage of such BASIS features as searching against hybrid databases of relational, bibliographic, and textual information.

4. **Overall Impression**

BASIS appears to have adequate ODBC and JDBC support, however neither driver is provided with the base system and at least in the case of the JDBC driver, the additional cost of the driver is significant. In addition, the JDBC driver has only been available for a short period as of the time of this writing, and so reliability of the driver may be of some concern.

3.7 Timing Requirements

(R) Timing Strings. The DOE-LSN shall meet the average response times shown in Table 5. The performance shall be achieved with 15 concurrent DOE-LSN users active on the system. [LSS-065]

The BASIS system cannot be evaluated at this time under the minimum required load of 15 users, or with the required 5 million pages of document data. However, the minimum timing requirements were analyzed with the smaller test collection of about 50,000 pages of document data.

These tests were all performed on a remote Windows NT 4.0 client machine. The client PC is a 450Mhz Pentium II running Netscape Communicator 4.5 to connect to the BASIS server. The machines are connected via a 10Mbit LAN. Since we are not testing BASIS under operational conditions, the load on the server is relatively low compared to its capabilities. The default web server provided with BASIS has three different methods of searching the document collection: the standard search form, a FQM command, and the 'search assistant' form. The FQM command entry interface was used in these tests.

Query	Average response time (seconds)
INJD-T3-Q1	0.6
TEJA-T3-Q2	0.8

Table 3: Timing for Retrieval of Results List

1. Retrieval of query results list. LSS2-065-2
 - The DOE-LSN requires that the query results list be retrieved in 45 seconds for UNLV test queries INJD-T3-Q1 and TEJA-T3-Q2. Each query was made five times as a broad concept query in BASIS, the time to retrieve each result list was measured, and the average computed for each query over all trials computed. Table 3 summarizes the BASIS average response time for each query.
 - **Overall Impression**
 These average times are for the retrieval of the first twenty relevant documents according to the BASIS system. Given the size of the current collection and the light load on the system, it is unknown if BASIS will satisfy this requirement.
2. Retrieval of header fields for document identified in query results list. LSS2-065-3
 - The current implementation of the LSS prototype collection in the BASIS system does not allow for separate retrieval of the header data for a single document identified in the query results list.
 - However, the default is to have the header fields included with each document retrieved from the query results list.
 - The web server also allows the retrieval of all of the document headers by using the `expand` option. The header information is not returned for an individual document but for all documents in the result set.
 - The time necessary to retrieve the header information is included in the document timing section above. We did not attempt to report the timings for the header retrievals alone, since the amount of time required to return this information is so small as to prevent reliable measurement.
 - It is believed that this requirement will be satisfied by the BASIS system.
3. Retrieval of text data for document identified in results list. LSS2-065-4
 - The LSN requires that the first page of text be retrieved in five seconds, and each subsequent page in one second.
 - From a sample of 10 documents retrieved from a query results list, the time to retrieve the entire document from the BASIS server was measured. These results are shown in Table 4. Note: The default BASIS web server does not return page aligned data, it returns up to a maximum of 32,000 characters per request to display the “next” page. Depending on the document content, layout, and hits on search terms, the time needed to display a page can vary significantly. In order to display a page aligned version of documents, this capability would have to be designed into the system, database structure, and web interface.
 - **Overall Impression**
 From Table 4 it is shown that BASIS retrieves approximately three pages per second on average. From the results of this experiment, it is believed that BASIS will satisfy this timing requirement.
4. Retrieval of image data for documents identified in results list. LSS2-065-5
 - The DOE-LSN requires the first page image to be retrieved in ten seconds, with each subsequent image retrieved in two seconds.
 - The default web server distributed with BASIS does not provide support for display of the page images. However, it appears that this capability could be incorporated

Document	Page Count	First Page Retrieval time (seconds)	Additional Pages Retrieval time (seconds)
1	14	0.80	0.58
2	93	0.75	0.48
3	162	0.69	0.14
4	65	0.76	0.27
5	65	0.71	0.15
6	97	0.71	0.47
7	64	0.78	0.45
8	78	0.97	0.81
9	19	0.70	0.32
10	7	0.72	0.32
Total:	664	0.76 (average)	0.37 (average)

Table 4: Timing for Retrieval of Document Text

into a custom web server without too much difficulty. We believe that the time required to access and display a page would be within the requirements.

- **Overall Impression**

A customized web server would have to be developed to allow the display of the page images. We believe that this interface could satisfy this requirement.

Requirement Identifier	Function/Event	Conditions	Response Time (15/50 concurrent users)
LSS2-065-2	Retrieval of query results list.	UNLV test query INJD-T3-Q1 or TEJA-T3-Q2* Database contains headers for at least 5 million pages of documents. A total of 10 documents found.	45 seconds/70 seconds
LSS2-065-3	Retrieval of header data for document identified in query results list.	Database contains headers for at least 5 million pages of documents.	5 seconds/8 seconds
LSS2-065-4	Retrieval of text data for document identified in query results list.	Database contains at least 5 million pages of documents.	First page: 5 seconds/8 seconds Each subsequent page: 1 second at Main Facility 2 seconds at supported sites
LSS2-065-5	Retrieval of image data for documents identified in query results list.	Database contains at least 5 million pages of documents.	First page: 10 seconds/15 seconds Each subsequent page: 2 seconds at site 3 seconds at other sites

Table 5: Response Time Requirements

EXHIBIT A:
Printed BASIS Document