

A Comparison of Thunderstone Taxis 2.6.930169407 with DOE/LSN Requirements

ISRI Staff

Technical Report 99-06
Information Science Research Institute
University of Nevada, Las Vegas

November 7, 1999

1 Introduction

The following document reviews Thunderstone's Taxis 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

Thunderstone - EPI, Inc., was founded in 1981 and is a privately held California corporation. From 1980 through 1995, most of Thunderstone's product licenses were embedded within OEM packages developed and sold by other organizations. The increased popularity of the Internet has raised Thunderstone's profile in large single-site applications like those at eBay, Novell, Advance Publications, Pactel, Associated Press, Ziff-Davis, and Corbis.

Thunderstone's list of products include: 1. *Taxis*, an integrated SQL RDBMS that queries full text as well as standard relational database fields. 2. *Webinator*, an HTML document indexing package. 3. *Metamorph*, the text retrieval search engine embedded in Taxis.

Thunderstone has issued more than 400,000 licenses to more than 1000 customers for applications such as: litigation support, competitive intelligence, help desk, document management, Internet search and publishing, and real-time message handling. Thunderstone claims to be a product oriented R&D company within the area of advanced Information Retrieval and Management.

The Taxis System software requirements include:

HTTP Server such as Apache, NCSA, Microsoft IIS, and Netscape
Frames capable browser such as Netscape or Microsoft Explorer

2 Collection Preparation

Taxis is a flexible system in that a web interface can be designed to meet most, if not all, of the LSN requirements. Collection preparation and interface design will play an important role in meeting these requirements.

Taxis includes two software tools that aid in system customization:

1. *Vortex* the Taxis' web script language.

2. *Webinator* a web crawler and Vortex based search interface.

Like most Relational Database Management Systems, the interface to Taxis must be designed by the application developer. If the application requirements are explicit, as they are with the LSN, careful planning during the design phase is crucial to the system's usability.

Webinator is a straight forward method for loading an existing HTML collection. It does not however, automatically load header data and the interface it produced satisfied few of the LSN requirements. For this reason, software for loading collections with header data was developed by ISRI staff using the Taxis API. The file format loaded into Taxis is shown in Figure 1.

3 Requirements Proficiency

3.1 General Requirements

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

1. Thunderstone states:

- Many people pursuing solutions to Year 2000 date problems desire compliance assurances from the products used in those software systems. This assurance is generally taking the form of a legalistic statement of compliance. The trouble with this approach is two-fold: (1) what is meant by compliance varies and/or is so all encompassing that no vendor can agree to the definition, and (2) lawyers are of no help if and when a system fails.
- There is no general software standard for the year 2000. As a result, many companies are generating an individual set of criteria and asking if Thunderstone products are compliant to their unique standard. One detailed summary of Year 2000 criteria is the *Millennium 2000 Program Document* from GTE Government Systems Corporation. Thunderstone has used this document as a standard by which to assess all of its current software packages.
- Thunderstone's packages have all been designed from the outset to have no internal issues with respect to the year 2000. However, as a database vendor Thunderstone's programs also rely on both the operating system on which our packages execute as well as the programmer who designs the application to not provide or display inaccurate date information to or from our products. In order to assess whether or not your installation of Thunderstone's packages are Y2K compliant, both the application and the operating system must also be considered.

2. Thunderstone explains **Date Issues:**

Binary Date Storage The native binary formats for Date and time information within our products are either 32 or 64 bit values representing the elapsed number of seconds since 00:00:00 UTC, January 1, 1970. This storage mechanism is not itself susceptible to problems with the year 2000 cross-over, but, the format will not operate correctly with dates on or after Jan 18, 2038. In our opinion this problem is actually far more prevalent in modern software than is the year 2000 problem. Almost all modern operating systems and 'C' programs use this format to manipulate date and time information.

Character to Internal Date Conversion Our programs contain several very powerful mechanisms for converting human readable date and time data into binary date storage format. These include:

- Parsing relative dates like "next Friday", "end of today", and "-1 month"
- Parsing formatted dates like "1994-03-05 06:30 pm"
- Conversion of combinations of individual form input fields

```

<html>
<head>
<TITLE>Preliminary Analysis of Geophysical Logs from the WT Series of Drill Holes, Yucca Mountain, Nye County, Nevada</TITLE>

<META NAME="docid" CONTENT="5069">
<META NAME="documenttype" CONTENT="Reports">
<META NAME="documentsubtype" CONTENT="Technical Reports">
<META NAME="publicationdate" CONTENT="19850000">
<META NAME="author" CONTENT="Muller, DC Kibler, JE">
<META NAME="keywords" CONTENT="Logging Geophysical Surveys Boreholes Data Handling Measuring Equipment And Systems Topopah Spring Member Stratigraphic Correlation Paintbrush Tuff Crater Flat Tuff Yucca Mountain Calico Hills">
<META NAME="pagecount" CONTENT="60">
</head>

<body bgcolor="#ffffff">
<h2>Preliminary Analysis of Geophysical Logs from the WT Series of Drill Holes, Yucca Mountain, Nye County, Nevada</h2>

<center>
<h2>Abstract</h2>
</center>

<p>
<Abstract>
  Geophysical logs from the WT series of drill holes correlate well with similar logs from other drill holes at Yucca Mountain, Nevada in the unsaturated zone through the same geologic units. The in-situ physical properties of the rocks from well logs are consistent with laboratory-measured physical properties of core from other drill holes. The Topopah Spring Member is concluded to have zones that are highly fractured and lithophysal in holes where the density and neutron logs are very "spiky" as noted in other cored drill holes. Low levels on the uranium trace from the spectral gamma-ray log indicate that fractures are neither healed nor filled with materials that concentrate uranium. Therefore, fracture permeability is expected to be high. This conclusion is consistent with fracture analysis from other drill holes on Yucca Mountain. The dielectric constant and dielectric resistivity logs correlate well with the epithermal neutron, borehole compensated density, and induction resistivity logs in the unsaturated zone.
</Abstract>
<hr>
<!-- Begin page 1 -->

```

[...]

Figure 1: HTML version of a document.

- Conversion of hexadecimal strings.

Testing of these functions within our software has been completed against all versions of our packages generated on or after September 2, 1998, and no issues are known to exist within these versions.

Conversion of Internal Date into Characters Our software packages must frequently convert their internal date and time format into humanly readable representations. These conversion functions are quite elaborate and will print dates and times in a great variety of ways. For the most part, display representation is controlled by the application programmer and not Thunderstone. It is possible for an application programmer to render confusing date information with our package (eg: "01/02/03") but this does not affect the accuracy of our internal calculation. Programmers are advised to correct any ambiguity when printing dates by including the century as part of the year.

Leap Year Calculations All currently shipping Thunderstone packages correctly account for leap years in conjunction with years before and after 2000.

Operating System Dependency The core efficacy of Thunderstone's date and time functions depend heavily on the operating system's native library functions. Most notably the 'C' strftime() library function, and to a lesser extent the entirety of the functions enumerated within the 'C' language "time.h" and "types.h" files. Any Year 2000 discrepancy within these operating system functions will be propagated through our software. The dependency precludes our making any Year 2000 statement of compliance in absence of prior certification by the operating system vendor on which our programs execute.

Operating systems known to have Year 2000 defects include: Microsoft Windows NT 4, Microsoft Windows 95, Microsoft Windows 3.1, Microsoft DOS, SGI IRIX 4.0.5 and before, and Linux versions 1.2.13 and before. Other operating systems may have flaws which remain undiscovered by Thunderstone.

3. Versions Tested and Warranty:

Upgrades to tested versions It would have required an unreasonable and extraordinary effort for us to have tested all versions of Thunderstone's software produced in the last 20 years of our existence. We chose to test only those packages shipping on or after September 2, 1998. While there may be no defect whatsoever in prior versions of our packages it is the customers responsibility to ensure that these versions will operate correctly within their environment. Customers may also choose to upgrade to the current version of any software package we currently publish.

Thunderstone will upgrade Customers who have purchased ongoing maintenance from Thunderstone for the cost of shipping and handling alone. Others may upgrade for the lesser of: 1% of the total license fee multiplied by the elapsed number of months to the date of upgrade or 15% of the total license fee.

Warranty The warranty for a Thunderstone product is fully described in the Software License Agreement that accompanies the product, and we recommend that customers read those warranties to understand their rights. The information we are providing here about Year 2000 readiness does not constitute an extension of any warranty for Thunderstone products. We are providing this information to assist customers to prepare for the year 2000 and beyond.

If you have further questions regarding Year 2000 compliance please contact the Help Desk at 216-631-8544.

(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. Taxis satisfies this requirement.
2. Considering the idea of accessibility on the Internet in the form of client/server communication, the Taxis FAQ specifically states that operation in “client/server mode...is the intended purpose of Taxis.”
3. Taxis includes the Webinator package, which provides a web-based interface that may be used by clients across the Internet. In addition, Taxis includes several tools for developers to construct custom applications that can make the database accessible to remote users across the Internet. The usual intention is for developers to build a web-based user interface, although non web-based interfaces are also a possibility.
4. There are factors that may affect accessibility of the system, all of which are considerations for any Internet communications, not just in the use of Taxis. These include:
 - 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.
 - 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 Taxis 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 Taxis, 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 and a response is received, many users may feel that the system is too slow to use.
5. **Overall Impression**

It is clear that accessibility of databases across the Internet is a major priority for Taxis. Tools are provided to enable the construction of applications that can accomplish this in several different ways.

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

1. Taxis satisfies this requirement.
2. Taxis is accessible by Windows-based clients in a number of ways; the methods of access depend on the user interface that has been developed.
3. The most widely usable interface would be web-based. Windows clients can easily make use of a web-based user interface through the use of a web browser (such as Netscape Communicator or Microsoft Internet Explorer).
4. Non web-based user interfaces to Taxis can also be built for the Windows platform using the supplied tools. For example, it should be possible to build a Visual Basic application that interacts with Taxis across the network via ODBC.
5. **Overall Impression**

Users of Windows systems can access Taxis in a number of different ways. How exactly this happens is dependent on the user interface that has been developed.

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

1. Taxis satisfies this requirement.

2. Taxis runs on at least the following platforms:

Operating System	Platform
AT&T SVR4	Intel x86
BSDI 4.x	Intel x86
Digital Unix 4.x	DEC Alpha
Hewlett Packard HP-UX 10.x and 11.x	HP-9000
IBM AIX 4.x	RS/6000
Linux	Intel x86
Microsoft Windows 95/98	Intel x86
Microsoft Windows NT	Intel x86
SCO 5.0	Intel x86
SCO Unixware	Intel x86
SGI IRIX 4.x	Silicon Graphics
SGI IRIX 5.x	Silicon Graphics
SGI IRIX 6.x	Silicon Graphics
Sun Solaris	SPARC and Intel x86
Sun Solaris 7 (64-bit)	UltraSPARC
Sun SunOS 4.x	SPARC

Note: Only Sun Solaris/UltraSPARC 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. Taxis returns a list of documents that meet the query criteria.
2. The user can sort results by document relevance or by selected fields in either ascending or descending order. The user can specify sort order within the SQL query language. An interface can be built with the Taxis API to simplify sort order selection.
3. Query for Document can be implemented as a basic function of Taxis. With *Webinator*, a simple point-and-click interface provided with Taxis, a user can enter query text and select simple query options including proximity and word forms. Webinator only allows search terms to be entered in natural language form; a complete SQL query using this interface is not supported. Figure 2 shows the Webinator query interface.
4. A web interface can be built on top of Taxis to query for document. Interface design would be a prerequisite for this system to meet LSN requirements.
5. **Overall Impression**

Taxis has the capability to return a list of documents that satisfy a given query in relevancy order or in ascending/descending order by selected header fields using SQL. A web interface can be built on top of Taxis to query for document. Interface design would be a prerequisite for this system to meet LSN requirements.

(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]

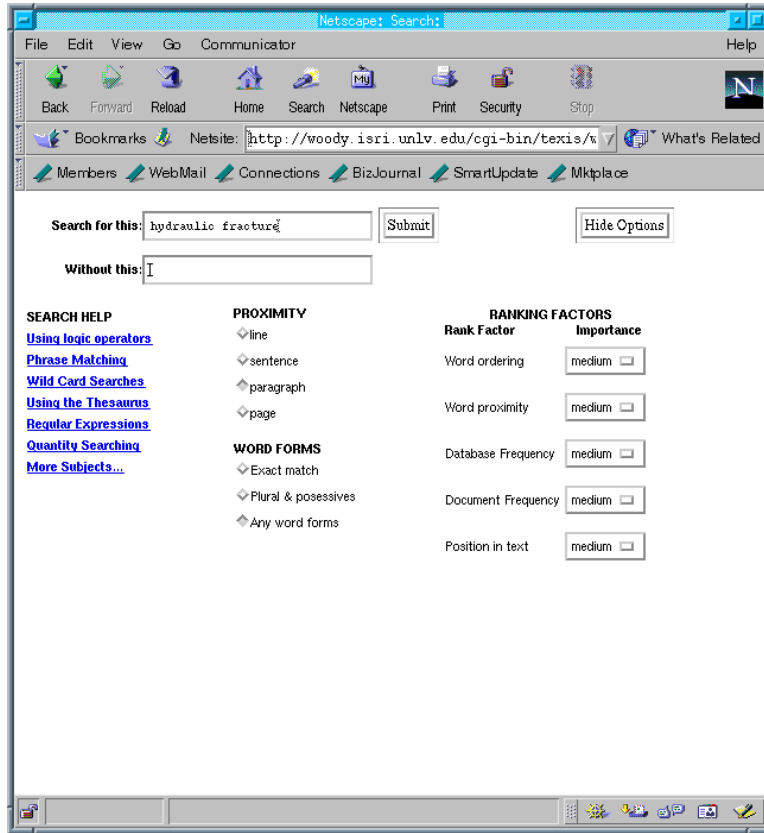


Figure 2: Taxis Webinator Query Interface

```
SELECT Title from HTML where Title like 'DOE' and
Body LIKEP 'Yucca Mountain' order by Title;
```

Figure 3: Header Field Querying with Taxis SQL

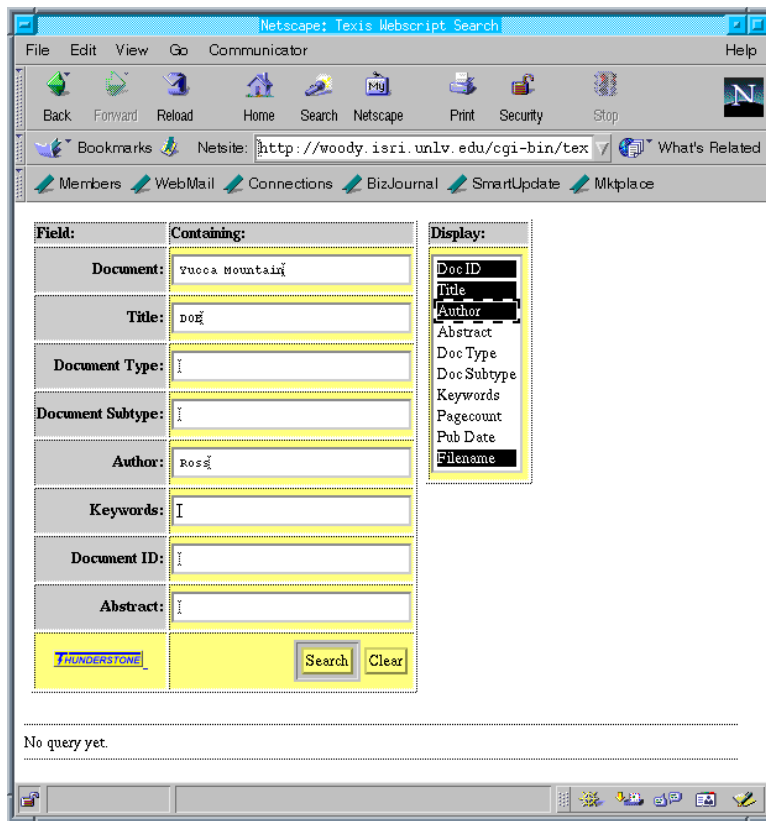


Figure 4: A Simplified Interface for Header Field Searches

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.

1. Taxis has the ability to return relevant documents when queried with specific header fields or on combinations of header fields by issuing the SELECT statement. An example fielded query is shown in Figure 3.
2. Using the Taxis API, an interface can be built to simplify header field searching. Figure 4 shows a simple interface designed for our evaluation.
3. SQL relational operators allow range value searching in date fields, etc.; equality is used by default.
4. Results retrieved can be ordered by relevance or in ascending or descending order as specified by the user.
5. **Overall Impression**

Taxis provides header field searching. A simplified interface is a prerequisite to make header field querying usable for an inexperienced user of the LSN.

- (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. Taxis provides the capability to specify character strings to search the full text of the document collection. Text searching is an integral part of the Taxis system.
2. The Taxis system query language is based on standard SQL syntax with extensions to the LIKE clause to query large text fields. The LIKE clause may include *Metamorph queries* for full text searching (See Requirement "Text Query Parameters" for details on Metamorph). For our query translation in this evaluation, we employed both SQL and Metamorph. The LIKE predicate has several forms in Taxis SQL:

LIKE as is customary in SQL, is used for *pattern matching*. The Taxis system extends the use of this operator by allowing all Metamorph queries to be substituted in the LIKE clause. Documents returned using the LIKE operator are not ranked by relevance.

LIKE* is used for full text searching but uses an internal formula to calculate document rank based on the presence or absence of terms in the document. There are several variables that control the ranking algorithm; these can be tuned by the system administrator.

LIKE* applies the same information for ranking as LIKER but also incorporates proximity of search terms into its calculation.

LIKE3 is an optimized form of LIKE in that no linear post-search of the text is done. This operator should only be applied when search requirements are narrow and speed is of the essence.

***Note:** LIKEP and LIKER must be used for ranked retrieval and must be applied to an indexed collection. LIKEP and LIKER queries may not include all Metamorph operators. See Table 1 for details.

3. The Webinator interface (shown in Figure 2) provided with Taxis allows Metamorph querying by applying the LIKEP operator. This allows the user to enter natural language queries but limits the use of Metamorph operators (see Note above).
4. Taxis lists several optimization techniques that will improve query response time. These should be evaluated by the system administrator to determine which index building techniques would prove most effective for the LSN's retrieval objectives.

@0 safety waste isolation +design +exploratory +shaft w/sentence "10 CFR 60"

Figure 5: Metamorph Query with some Query Language Operators

<i>Texis Query Language Operators</i>		
Operator	Usage	Effect
Mandatory Inclusion	+<term>	Documents returned <i>must</i> include any query term prefixed with +.
Mandatory Exclusion	-<term>	Documents returned <i>must not</i> include any query term prefixed with -.
Intersection of Terms/OR ¹	@<number_of_intersections>	Intersection is applied to a set of terms, where no intersections (@0) is equivalent to OR. The maximum number of intersections (number_of_terms-1) is equivalent to AND.
Thesaurus Expansion	~<term>	Enables or disables thesaurus expansion. See a complete discussion of thesaurus expansion in the Query Assistance section.
Character Wildcarding	* (asterisk)	Matches up to 80 characters in the position indicated. More than one * may be used.
Regular Expressions ¹	/ (forward slash)	Invokes REX (R egular E Xpression), allowing searches for any fixed or variable length regular expression. REX would include single character wildcard searching.
Within	w/<document_segment>	The w/ (within) operator is used to designate that search terms should appear within the specified segment. The segment of text can be system defined segments (line, sentence, paragraph, page, document), administrator defined segments (using start and end tags), or a specified number of characters.

Table 1: Query Operators in Taxis Metamorph

5. Overall Impression

A simplified interface that employs both SQL and the Metamorph query language to its fullest should be designed for the Taxis system. Interface design is typically expected for applications in relational database systems.

(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]

1. Taxis' Metamorph, the text search engine, provides these query capabilities using the operators described in Table 1 and through parameters set by the system administrator. These operators can only be used within a Metamorph query. Figure 5 is an example of a Metamorph query. Note that a Metamorph query is the clause that follows the LIKE operator in a standard SQL query.
2. These capabilities are an integral part of the Taxis system.
3. Taxis employs logical or pre-marked segments for proximity searching except in the case of character count as specified by the user. This makes sense as long as the method for defining logical segments fits the document collection's format. With OCR text, this

¹These operators cannot be used in combination with relevancy ranked retrieval (i.e., LIKER, LIKEP).

may not always be the case. Segment definitions though can be re-defined and new segments can be marked by the system administrator that would best suit the collection and its users.

4. Taxis accommodates root searching as an element of its *Morpheme Processing*. This process uses several known stemming techniques to derive a term's root.
5. Taxis does not provides the ability to combine the results of one or more previous searches with a new search.
6. **Overall Impression**

Taxis's text query system, Metamorph, provides a set of operators for searching full text. It's syntax is cryptic and would require a more usable interface for Metamorph to be a useful tool for searching the LSN document collection.

(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]

1. Taxis is capable of querying header fields and document text in a single query.
2. As indicated, the Taxis SQL syntax would be considered too complex for general use so a suitable interface should be designed to provide a simplified means of querying multiple header fields and document text in a single query.

3. **Overall Impression**

Since header field names must be specified in the text of the query, the application programmer should design a consistent and simplified interface for combining header and text field searches.

(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. Taxis does not provide any specific query status indicator while the system is searching other than what is displayed by the browser.
2. 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.
3. 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.
4. If the user is familiar with the browser in use, then the browser indicator and the "stop" icon can aid in query status and query termination.
5. Taxis provides the total number of documents retrieved for a query.

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.

1. Taxis satisfies this requirement by including the following capabilities:

Natural Language Taxis is a relational database that extends the LIKE operator for full text querying. Any natural language or *Metamorph* text query can follow the LIKE operator. An interface should be built for text field queries so that the SELECT clause is understood.

Regular Expressions Taxis includes a regular expression utility, REX, that provides the ability to search for regular expressions. Types of expressions that can be found using REX include pattern strings such as phone numbers, formulas and dates.

Approximate Pattern Matching XPM is a tool included with the Metamorph query language that allows approximate pattern matching. A percentage of proximity to the entered pattern is specified to locate “close” strings in the text.

Numeric Pattern Matcher NPM matches numeric patterns in document text.

Thesaurus Taxis includes a 250,000 word association thesaurus. Word associations are stored in an *Equivalence File*. The included thesaurus can be augmented by creating a *User Equivalence Files*. These files are comprised of word list associations where part of speech (e.g. noun, pronoun, verb) can be included.

Morpheme Processing *Morpheme Processing* is the Taxis system’s equivalent of root searching. It applies prefix and suffix removal as well as stemming to find root words for searching. These processing steps are under the complete control of the system administrator by setting parameters and modifying predefined word lists.

2. Overall Impression

The Taxis system relies heavily on pattern matching techniques to extend and improve the types of searches that a user can perform. These kinds of queries are very powerful but are also very complex to construct. Without a well-developed interface, the typical user would probably not employ and therefore, would not benefit from this kind of query assistance.

3.3 Display Capabilities [LSS2-012]

NOTE: All display capabilities are dependent on the user interface. Taxis provides users with a simple interface called *Webinator*. To evaluate and test certain requirements, ISRI built other interfaces. Responses to the following requirements are based on *Webinator*, the interfaces built for this evaluation and *Webinator*, *Taxis*, and *Metamorph* documentation.

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

1. Taxis provides the ability for a user to select a document to be displayed from the results list of a query. The default interface provided with Taxis is shown in Figure 2.
2. A document can only be viewed after a query has been run. No documentation was found that will allow the user the ability to view a document without running a query first. The document can be viewed by selecting one of the document types listed below from the results list in *Webinator* (See Figure 6).
3. The *Webinator* interface provides two viewable versions of a document – *HTML* and *Match Info* format:

HTML The pages of each document have been combined to make a single HTML page. Page links are included for random page access. Thumbnail TIFF images for each page are embedded in the HTML with its corresponding text. Refer to Section 2 for a description of the way this collection was prepared.

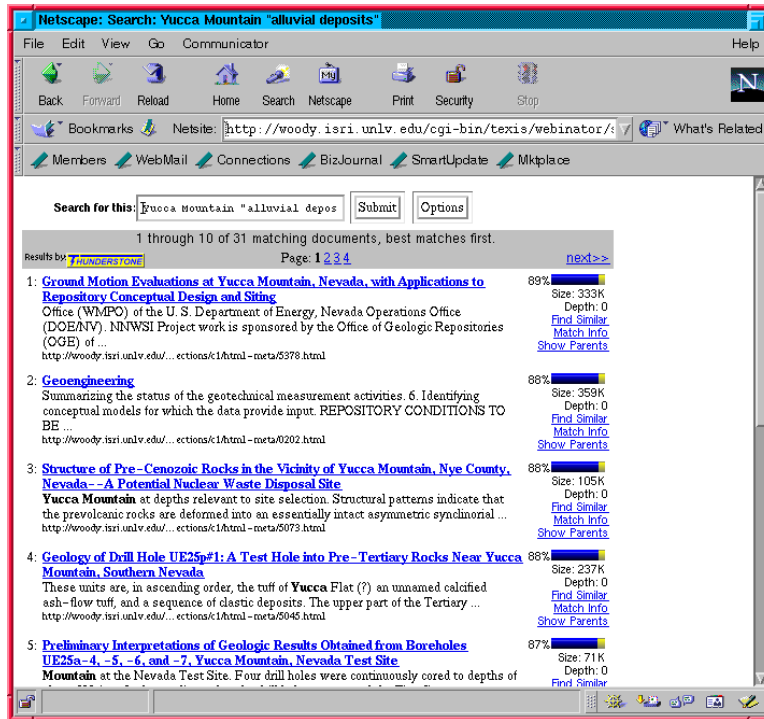


Figure 6: Webinator Results List

Match Info The document is again presented as a single page. Search terms are highlighted and the interface allows for jumps to the subsequent search terms. Images are omitted.

4. Overall Impression:

Texis satisfies the document display requirement.

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

1. Header field display is not inherent in Texis and is not implemented in Webinator but can be achieved by modifying the user interface and the document collection.
2. In ISRI's modified interface, the user is able to select fields to be displayed before a query is run. Figure 7 shows the ability to display selected fields in the results list.

3. Overall Impression

To make Texis compliant with this requirement, a user interface was built to search and display particular header fields. These can be selected before the query is run.

(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. Texis meets this requirement in Webinator through either the HTML or Match Info formats.

HTML displays the complete document text that has been segmented into pages. Links to all pages are added for convenient random access.

Match Info displays the document provided by Texis with no page delimiters.

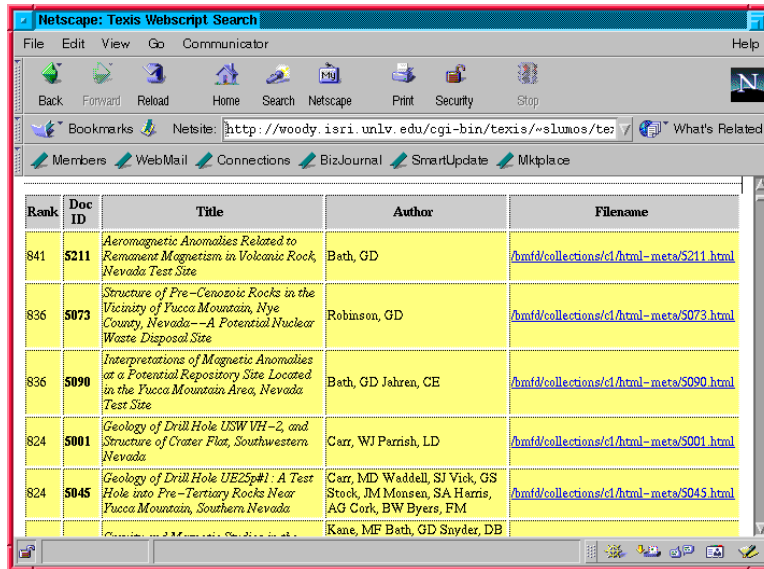


Figure 7: Header Field Display in Taxis

2. Overall Impression

Taxis satisfies this requirement with HTML mark-up added to the collection by ISRI. Other implementations of text page viewing may be possible by modifying the user interface.

- (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. Taxis provides this capability through Webinator with the Match Info display option.
2. Match Info shows the entire document with all search terms highlighted. The ability to jump forward to the next term is provided by linking terms. Figure 8 shows this ability in Taxis. This feature allows immediate location of the terms used in the query.

3. Overall Impression

Taxis provides term highlighting and the ability to move forward from hit to hit 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. Taxis does not directly satisfy this requirement. Only through the use of a web browser and an appropriate plugin can Taxis satisfy this requirement. The plugin selected should support TIFF image format.
2. The user can only view the image of the document after selecting a document from the results list.

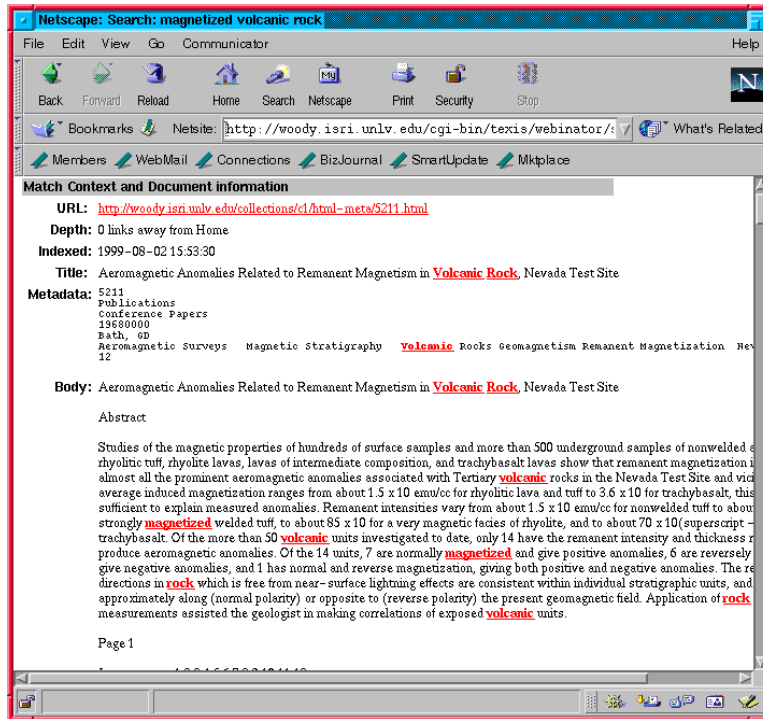


Figure 8: Highlighting Search Terms in Taxis

3. None of the images in our test collection were JPEG images but JPEG viewing is accommodated by most Internet browsers in use.
4. When viewing the oversized images the plug-ins automatically fit the entire image within the viewing area of the browser.
5. **Overall Impression**

A TIFF plugin viewer integrated into the browser will satisfy this requirement. Most plug-ins are easy to use and their features are self-explanatory. 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 Taxis satisfy this requirement.
2. For this evaluation, ISRI installed *AlternaTiff* a TIFF viewer plugin to meet this requirement. This plugin is equipped with buttons which perform the image augmentations listed.
3. **Overall Impression**

Taxis meets 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. Taxis can satisfy this requirement if 1) A plugin is installed which launches a new window for TIFF image display, or 2) Taxis' 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, including the one used in this evaluation, 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. Webinator and the interface designed by ISRI do not satisfy all display combinations listed in this requirement.
2. To view the header fields of a document the user must select these fields before running a query using the ISRI interface. The contents of these fields are listed as part of the results list.
3. The TIFF image for each document page is included as part of the HTML source. The user can select the thumbnail to bring up the image of the current text page being viewed. The images were incorporated into the documents during collection preparation (See Section 2 for details) by ISRI. This is not an inherent capability of Taxis.
4. The text of the document can be displayed only after a query is run.
5. Neither Webinator nor the interface designed by ISRI display the combinations: header and text, images and header, images and text in a single browser window. These can be implemented using Taxis' web script language called *Vortex*.

6. Overall Impression

Taxis does not inherently meet this requirement but does provide the ability to satisfy it. The implementation of this requirement will require some amount of programming time.

3.4 Printing Requirements [LSS2-013]

NOTE:

The printing capabilities of Taxis depend entirely on the browser in use. In addition, the functionality of printing and the appearance of the printed material depends on the interface design.

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. Taxis satisfies this requirement through the use of a web browser. No special capabilities of the system are employed.
2. Exhibit A shows a printed HTML document from the Taxis system. The format of the document was implemented by ISRI. It includes all document pages and image links. Other document components, like image pages, would have to be printed separately.
3. Exhibit A also shows a printout from the Match Info option of the Webinator interface. Search terms are highlighted.
4. **Overall Impression**
Taxis 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).

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

1. Printing a document header is not inherent in Taxis. ISRI included this capability when designing an optional interface.
2. The amount of detail contained in a particular header field can be selected before a query is run. This selection determines what header fields appear in the results list. With the ISRI user interface, the header can be printed as part of the results list but not as a separate unit.

3. Overall Impression

The header is associated with the results list as opposed to a selectable element for viewing. This feature though could be incorporated into the interface as a viewable and therefore, printable unit.

(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. Taxis can print all the text of a document, but there is no concept of “pages” that corresponds directly with the printed pages of an LSN document.
2. Since Taxis depends entirely on the browser for printing, only an “HTML page” can be printed which in this case, is the entire document. A range of pages can be printed in some browsers but they will not be equivalent to the pages in the document.

3. Overall Impression

Taxis is able to print a document in its entirety. Printing single pages or 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. Taxis does not supply a report generation facility.
2. It may be possible to implement such facility by adding columns to the system and collection tables reflecting the type of information required for reporting.

3. Overall Impression

Although Taxis does not provide a report generation facility, it could be implemented. Any report generation should be addressed during initial design and collection preparation.

(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. Taxis depends on the browsers capabilities as well as an appropriate plug-in to satisfy this requirement. This capability is not an integral part of Taxis.
2. Printing images, in particular TIFF images, may not be an inherent component of the default browser in use. In many cases, the user will need to download a plugin before they can view or print a standard image.
3. 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.

4. The ability to print a *selected range* or *all* page images of a document is usually not a feature of viewer plug-ins and was not built into the structure of the HTML documents in our test collection.
 5. **Overall Impression**
 Apart from modifying Taxis's interface, a plug-in 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 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. Taxis depends on the viewer plugin and printer capabilities to satisfy this requirement. This capability is not an integral part of Taxis.
 2. Printing an oversized image depends on the capabilities of the printer being used.
 3. **Overall Impression**
 Since all printing in Taxis is browser and plugin dependent, printing oversized images depends on 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. Taxis satisfies the ability to print the results.
 2. Webinator lists results in groups of 10 and this is the number of results which can be printed at one time. A summary of each document is provided in the result list.
 3. The ISRI interface can only print the results list in its entirety. The field information printed is determined by the user before the query is run. Figure 7 shows the results list with Rank, Title and Author selected for display.
 4. **Overall Impression**
 Depending on interface and needs printed result lists can vary. Since the interface is customizable, any desired information can be included in the result list 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, Taxis 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**
 Taxis' 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 Taxis is customizable, "Request Paper Copy" could be integrated into the interface design using Vortex and API.
 3. **Overall Impression**
 Taxis is customizable and an interface can be created to service specific needs. "Request Paper Copy" is one of the features required by the DOE that could be implemented using the system's programming language, Vortex, and the Taxis API.

- (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 could be integrated into a user interface in a way similar to its counterpart, “Request Paper Copy” using Vortex and the Taxis API.

3. **Overall Impression**

While this is not a standard feature, Taxis provides a programming language, Vortex, and an API for its implementation for a customized interface.

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 Taxis system.
2. The Taxis product allows the administrative user to monitor system components through the standard Unix process control facilities. Additionally, Taxis provides a program, `monitor`, that allows the administrator to monitor many Taxis system functions. The `monitor` process can be run in interactive or logging mode.
3. Additionally, the database administrator can use Metamorph to query and set a variety of system variables affecting performance and operation of the search engine.

4. **Overall Impression**

While Taxis does not supply a wide range of monitoring utilities, the combination of Metamorph, `monitor`, and Unix process control make it possible to monitor and control the Taxis system.

- (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 marginally satisfied by the Taxis system.
2. Taxis does not provide any tools specifically designed for monitoring user session activity or queries.
3. Taxis does provide *triggers* that allow the administrator to be automatically notified of specific events.
4. These triggers allow an arbitrary Unix command to be executed when specified events occur.

5. **Overall Impression**

While no specific user level or session administration tools are provided with Taxis, through the combination of the Vortex web API and the system triggers, more extensive system and user level monitoring tools could be developed to satisfy this requirement.

- (R) **Database Administration Tools.** DOE-LSN shall provide authorized users the capability to assess the availability, integrity, and performance of the databases of 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. This requirement is marginally satisfied by the Taxis system.
2. Taxis does not specifically address this requirement in the suite of administration tools provided with the system.
3. Database *triggers* and the *monitor* program can be used to monitor some system parameters.
4. Custom administrative tools would have to be built to fully satisfy this requirement.

5. **Overall Impression**

The Taxis system does not provide pre-built tools that satisfy most of the administration requirements. However, the combination of Vortex, command line tools, and various APIs provided with the system would allow a developer to create very powerful administrative tools. These tools, if designed and implemented properly, would satisfy the LSN requirements.

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. Taxis satisfies this requirement.
2. Querying and retrieving results through a web server is the usual method of interacting with Taxis. Taxis provides three major facilities for interfacing with a web server: the Webinator package, the Taxis Webscript environment, and an API for constructing custom applications.
3. The following description was taken from the Webinator online documentation:

Webinator is a web walking and indexing package that will allow a website administrator to create and provide a high quality retrieval interface to collections of HTML documents. Webinator serves as an example of the type of applications that can be built around Thunderstone's Taxis RDBMS and Webscript.

4. The Taxis Webscript environment is included with the system. It is an integration of Taxis (the RDBMS), Metamorph (the text retrieval facility), and Vortex (a CGI-scripting language and compiler). The Vortex language is actually an extended version of HTML, having certain programming features. Upon receipt of a client's request for a Vortex file, the web server will invoke the Vortex CGI compiler, which then handles the execution of the Vortex file. Typical execution of such a file might include sending a query to Taxis/Metamorph, retrieval of the result, and the formatting of the result for the client to view.
5. Custom applications to handle communication between Taxis and a web server can be built using the supplied C API. Such applications would normally function as CGI programs, invoked by the web server upon receipt of a query. The application would call the Taxis API routines to execute the query and retrieve results, then handle the processing of those results and formulating a response to the client. There are example programs included that demonstrate the use of the API. It should be noted that although the Taxis API is C-based, many other programming languages provide facilities for linking with such libraries, so developers should not be limited to programming in C when using the API.

6. **Overall Impression**

Accessing Taxis through a web server is clearly the most emphasized method of interacting with a database. Taxis provides a complete pre-built web-based user interface, a scripting environment to quickly and easily build custom web-based interfaces, and an API to enable developers to build their own CGI programs from the ground up.

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

1. Taxis satisfies this requirement.
2. Taxis provides a C API that allows developers to build programs that obey the CGI standard and handle communication between Taxis and a web server. Whether or not such programs conform to the CGI standard is entirely up to the developer.
3. Also note that the program that handles the processing of Vortex files is invoked by the web server as a CGI program.
4. Although no Perl5 module to enable access to Taxis is included, and no third party module could be located at the time of this writing, it should be possible to construct such a module given the Taxis C API.

5. **Overall Impression**

Taxis provides the ability to build CGI programs to perform virtually any type of processing of queries or results that is desired. Access to Taxis from the Perl5 language is not provided, however construction of a facility to provide such access should be possible using the Taxis API.

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

1. Taxis at least partially satisfies this requirement.
2. Taxis provides ODBC support as described in the following quote, although the Taxis API documentation indicates that the API is the preferred programmatic interface.

We provide an ODBC driver for Windows that will allow you to quickly create your client interfaces in products like Access, and Visual Basic. The ODBC driver is currently only available to talk to 32-bit servers, and requires TCP/IP.

3. Only one mention of JDBC support was found in the documentation. It stated that JDBC support was added to Taxis in 1996. No other information, Java classes, or example code was found.

4. **Overall Impression**

Taxis seems to have adequate support for ODBC connectivity, however JDBC support is unclear.

3.7 Timing Requirements

(R) Timing Strings. The DOE-LSN shall meet the average response times shown in Table 4. The performance shall be achieved with 15 concurrent DOE-LSN users active on the system. [LSS-065]

The Taxis 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 Taxis server. The machines are connected via a 10Mbit LAN. Since we are not testing Taxis under operational conditions, the load on the server is relatively low compared to its capabilities.

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, the time to retrieve each result list was measured, and the average computed for each query over all trials computed. Table 2 summarizes the Taxis average response time for each query.

Query	Average response time (seconds)
INJD-T3-Q1	9.0
TEJA-T3-Q2	32.2

Table 2: Timing for Retrieval of Results List

Document	Page Count	Retrieval time (seconds)
1	44	2.0
2	38	2.5
3	25	1.5
4	55	3.0
5	16	1.0
6	195	10.0
7	69	4.5
8	50	2.5
9	126	8.0
10	30	2.0
Total:	648	37.0

Table 3: Timing for Retrieval of Document Text

- **Overall Impression**

These average times are for the retrieval of all documents considered relevant by the Taxis system. Note that Taxis returned far more relevant documents for TEJA-T3-Q2 than INJD-T3-Q1, and this accounts partially for the disparity in retrieval time. Under these testing conditions, it is unknown if Taxis 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 Taxis system does not allow for retrieval of the header data for a document identified in the query results list.
 - However, it is possible to have the header fields included in the query results list.
 - This substantially increases the size of the query results list, but only increased the average response time for INJD-T3-Q1 to 44 seconds.
 - It is believed that this requirement will be satisfied by the Taxis 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. Under our implementation of the LSS prototype collection in the Taxis system, the entire document is returned when selected from the query results list. Note that thumbnail images for each page in the document are returned as well as the corresponding text.
 - From a sample of 10 documents retrieved from a query results list, the time to retrieve the entire document from the Taxis server was measured. These results are shown in Table 3.
 - **Overall Impression**
From Table 3 we can infer that Taxis retrieves approximately 18 pages per second. From the results of this experiment, it is believed that Taxis 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 current implementation of the test collection in Taxis does not allow for retrieval of the entire collection of page images for a document. A sample of ten pages from a

document were retrieved. The average retrieval and display time was approximately three seconds per page.

- **Overall Impression**

From this experiment, it is believed that Taxis will satisfy the timing requirement for retrieval of the first page. It is unknown if the timing requirement for each subsequent page will be satisfied.

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 4: Response Time Requirements

EXHIBIT A:
Printed Taxis Document