Texis FAQ

Thunderstone Software

April 15, 2024

Contents

Chapter 1

General

1.1 How is Texis different from other search engines?

Texis is the only search engine with the structure of a SQL relational database (rdbms). SQL as used here means Structured Query Language – not Microsoft's product named with that term! SQL is an industry standard defined by the American National Standards Institute (ANSI), and its counterpart, the International Organization for Standardization (ISO). All major database vendors use SQL as their query language.

SQL provides many advantages for addressing complicated search requirements. It also provides you with the confidence of a reliable, well-defined path for implementing unanticipated new search functionality in the future. SQL is a rich, mature, open standard used by hundreds of thousands of database application developers around the world.

All other search engines provide a much narrower range of capabilities based on proprietary interfaces. No other search engine provides the versatility of using SQL as its application development model.

1.2 How is Texis different from other relational databases?

Texis is the only relational database that can store and search text documents of unlimited size within standard database tables. All other solutions that purport to accomplish this employ, either explicitly or "under the covers," a loosely coupled external text index, and store documents in a binary large object (blob) field. That approach causes major bottlenecks.

1.3 What's so hard about integrating text-search with an RDMBS?

Text-searching and relational database management are radically different paradigms for organizing and retrieving information. They were developed over decades as completely separate technologies and do not marry easily. Thunderstone has devoted more than 10 years to solving this problem; it is our "core competency." Thunderstone is the only software vendor to have undertaken and solved this challenge head-on.

1.4 Is Texis a content management system?

Many of our customers use Texis for content management. Examples include legal document archiving and editorial publishing systems. However, Texis "out of the box" is not a finished content management application. You must customize it for that purpose – see below regarding application development. One of the benefits of using Texis for content management is that you can make it do what you want. If off-the-shelf content management solutions don't quite do everything you need, you'll end up bringing in consultants (programmers) to customize it anyway. Sometimes it is preferable to start with a more generic platform, and design exactly what you need from the start.

1.5 Is Texis an e-commerce system?

Many e-commerce web sites use Texis. Some are powered entirely by Texis technology. Others use Texis as a search engine, together with commerce tools of other vendors. Texis is an ideal application development platform for various e-commerce applications. Thunderstone also provides some generic commerce application scripts that may be customized with the tools discussed below.

1.6 Is Texis a portal system?

The word "portal" has come to mean a uniform web "front-end" to disparate "back-end" computer systems. Texis includes many portal-building features. The most important relate to web fetching, i.e., retrieving content from other web applications in real-time. A typical use of this ability is to create a "federated" search application that integrates results from disparate sources. Texis actually is a whole suite of tools that together enable a developer to create, deploy, and maintain web-based applications. The tool suite encompasses: a scripting environment; data importing utilities, a web crawler; a database; a search engine; and even a web server.

1.7 Is Texis a knowledge management (KM) system?

Knowledge Management is a broad term for sharing of knowledge within an organization systematically, instead of informally as is the human tradition. Tools for searching archives are fundamental to KM. In many cases, an effective internal search engine will give you the greatest KM benefit for the least investment. Texis is designed to meet the most sophisticated internal search needs. The idea of KM may encompass a variety of other applications such as cataloging of individuals' expertise. Texis provides tools for creating such applications, but they are not "out of the box" features.

1.8 Database X has a full-text-search feature – How is that different from what Texis does?

Database X in fact has tied in a separate text-indexing module by means of a "foreign-key join," an extremely inefficient technique. It is suitable only for small databases and light user loads. And text searches will be slow!

1.9 Search engine Y can index a relational database – How is that different from what Texis does?

Search engine Y ignores SQL logic and treats database records as documents. So you lose ability to do sorting on other columns, real-time updates, and many other capabilities that are inherent in a relational database.

1.10 Search engine Z can communicate in SQL – How is that different from what Texis does?

Search engine Z uses only a subset of the SQL language, and does not perform some of the basic functions of a database. For example, it probably doesn't support JOIN operations, the DISTINCT argument, the `GROUP BY` command, the `HAVING` clause. These are all generic features of the SQL language. And when records are updated (changed), it must rebuild the search index in a time-consuming batch operation. In between index rebuilds – which may occur only once a day – searches match the old stale data, not the new data. A true relational database (of which Texis is one) never allows the indexes to get "out of sync" with the data.

1.11 What is the difference between Texis and Thunderstone's other products Texis Web Script, Vortex, Webinator, and the Search Appliance?

Texis is the core technology, encompassing the database and search engine capabilities.

Texis Web Script also is known as Vortex. This is an application development (scripting) tool set. It is bundled with Texis. Texis Web Script is a superset of HTML, with extensions for passing calls to the Texis database. Returned data is dynamically marked up, normally with HTML but optionally with JavaScript or any other markup language. Texis applications may alternately be created with a variety of other technologies.

Webinator is an application of Texis. The Webinator crawler (dowalk) extracts the content of web pages and stores each as a record in a Texis database. The database thus becomes a search engine for the spidered pages. It can be queried through normal Texis SQL `SELECT... LIKE` statements, but a generic web interface also is provided. Webinator is given out as an example Texis application, but under conditions defined by its acceptable-use policy (see license addendum), it may be used for indexing a web site(s) without purchasing a full Texis license.

The Search Appliance also is based on Texis. It encompasses the features of Webinator, with some additional capabilities such as the ability to crawl file systems and respect document permission settings. The Appliance is created to be a "turn-key" solution, so that the customer does not need to install or configure software or operating system features. The Search Appliance is designed to be administered by a business user, as opposed to a technical user.

1.12 What is the "target market" for Texis?

Target applications include online publishing, interactive catalogs, classified advertising, digital asset management, intelligence, and of course, web searching. What these all have in common is that they require both structured and unstructured types of searching. For example, product catalogs typically contain unstructured text (name, description, etc.), as well as structured content (size, price, inventory number, etc.). Users may wish to search by description; or navigate by price range; or both in combination. Texis is the premier solution for providing text searching tightly integrated with traditional structured database querying.

Chapter 2

Texis as a Database

2.1 Which database (RDBMS) does Texis use?

Texis does not "use" another database; it is a complete database itself. However, it can be used as a search engine for content residing in any other database.

2.2 Is the Texis database proprietary?

Texis is non-proprietary in the sense that is integrates the ANSI SQL relational database standard. Texis is proprietary in the same sense that Oracle is a proprietary database. But see below about the `LIKE` clause syntax.

2.3 Does Texis have stored procedures?

Yes. Procedures are stored in server-side compiled scripts.

2.4 Does Texis do joins?

Yes. The technique of joining two or more tables together in a relationship is what makes a database "relational." As with any database, in designing for scalability, care must be taken to avoid excessive reliance on joins, which tend to be resource-intensive.

2.5 How are search-engine type queries expressed in SQL?

In Texis, full-text search terms go inside the `LIKE` clause. An expanded syntax, including the familiar + and - operators, is supported. This expanded syntax by necessity is implemented as an extension of SQL simply because traditional search-engine type queries are not defined in standard SQL.

2.6 What is a Metamorph index?

This is Thunderstone's name for an inverted (search-engine) index on a column of unstructured text, as distinguished from a b-tree (sorted-order) index normally used on numeric or string database fields. This following is somewhat technical and is intended for experienced database programmers.

A Metamorph index is set up and used in a manner analogous to any other database index:

```
create metamorph index descriptionindex on products(description);
```

As an illustration of the power this provides, consider a typical Texis query of this model:

```
SELECT id, name FROM products WHERE description LIKE 'big fancy gizmo'
   ORDER BY price;
```

The Texis database optimizer uses the metamorph index on the description field, along with the B-tree index on the price field, to quickly and efficiently resolve this query. (From a more technical point of view, you'd probably create a single compound index combining the characteristics of those two indexes, for even better performance.)

Any other database, to accomplish something similar, would receive no help from the database optimizer. The text index is a black box to it. It can only hand off the gizmo query to a separate text index; then create a temporary table containing the text-search results; then do a join between that and the table containing the price information.

Texis is the only relational database that can resolve a query of this type without a join. This makes Texis many times more efficient.

2.7 Are documents stored within Texis, or as separate files?

Either! It depends on the circumstances. Web-searching is a typical example of indexing external documents: the Texis crawler extracts information about the pages and builds an index (database) based on that; search results consist of links to those pages. On the other hand, in an auction application, the original information typically exists entirely within the database: users input their listings directly into the database; and search results consist of links to records within the database.

2.8 Do I need a Texis DBA (database administrator)?

Probably not. Administering a Texis database is simpler than administering other databases such as Oracle. Texis runs as an application on top of the operating system and does not usurp operating system functions. So backing up a Texis database, for example, can be accomplished simply by copying a directory (but see below regarding redundancy). Texis does have various administrative aspects and configuration options, but overseeing those usually is handled by the application developer(s).

2.9 Can Texis handle BLOBs (binary large objects)?

Yes. Texis has a blob-type field useful for storing graphics or other binary data. But note that in Texis, textual content of any size usually is put in a variable-size varchar field. This provides superior text-indexing and searching functionality compared to storing text into blobs. But if you have binary content, Texis can manage the storage of files much more efficiently than an OS file system! That is because Texis keeps track of each record's location on disk, and can fetch it with a single disk seek-and-read operation; whereas operating systems are un-indexed, so that fetching files typically takes four or more seek-and-reads to search through the directory structure.

2.10 Does Texis do data mining?

Yes, definitely. The idea of data mining carries the connotation of a large archive of data collected from other business processes. The archive is used for analysis, typically in a search for trends or relationships not obvious in the normal course of business. Texis is an excellent platform for data mining and may be superior to traditional databases for this purpose, depending on the kind of data and analyses needed. Texis is most versatile for querying data that combines structured and unstructured elements. Many business

databases containing unstructured text fields, such as customer correspondence or support logs. Insights may be found in patterns either in the structured or unstructured data, or in a combination. Texis is the market's leading tool for such research.

Chapter 3

Texis Search Technology

3.1 Does Texis handle natural language queries?

Yes. Users may enter any natural language question. By default, matching records or documents are presented in relevance rank order. There are many settings for "tuning" the rankings.

3.2 Can Texis index PDFs, word-processing documents, or other formats?

Yes. Texis can index most common document formats. It will also extract and text in binary files, such as a photo containing a caption in ASCII.

3.3 Does Texis highlight "hits" (word matches) in the results?

Yes. Texis can even pass the appropriate information to Adobe Acrobat to perform the highlighting within a PDF document. Hit highlighting is completely customizable with CSS classes and stylesheets.

3.4 Does Texis create a summary of the each result item?

Yes, we call this an abstract. Normally it is centered around the most representative "hit" words in the results, and the size of the abstract may be specified.

3.5 Does Texis handle phrases? Wildcards?

Yes, both. A typical search form will consider text within quote marks as a phrase, and the asterisk character as a wildcard. If desired, Texis will accept wildcards within or at the beginning of a word, as well as at the end. These features are under the control of the application developer, who may turn them on or off, or change their behavior in various ways.

3.6 Does Texis support Boolean logic?

Yes. Full Boolean logic is standard within the SQL language. Texis also understands the + and - operators popularized by web search engines. And Texis understands set logic, which can be used to express a command of the style "Find records containing n or more words of my query." Absent explicit operators, the default logic is specified by the application developer.

3.7 Does Texis have fuzzy logic?

Yes. The Texis facility for accomplishing this is called approximate pattern matching. This generates a similarity measure between any two words or patterns, expressed as a percentage of closeness. The user or application developer may control the degree of closeness. This capability most commonly is desired to accommodate spelling mistakes in either the queries or the data. It can be useful in searching scanned documents, which tend to have errors resulting from the imperfect OCR process. Developers should use this feature with caution, however. Fuzzy logic, by its nature, brings back some records unrelated to the either the user's query words or the intended meaning. This tends to confuse and annoy users not expecting this style of response.

3.8 Can Texis index documents stored on multiple servers?

Yes, elementary! Texis may create a searchable index of documents anywhere on a network or on the Internet.

3.9 Can Texis sort results by date (or by price, or rating, or whatever)?

Yes. Texis's sorting power is one of its most popular features. You may sort the results of a text search by any field in your data. For example, if your database contains an `author` field, you can sort search results by author. This works efficiently even on large result sets, by taking advantage of the powerful sorting capability inherent within relational database technology. Texis can quickly sort tens of thousands of hits or more. Other search engines either bog down sorting more than a few hundred items, or else their sorting capabilities are much more limited. For example, one major search engine cannot perform relevance-ranking together with sorting; another can sort by date only, not by other fields.

3.10 Can Texis organize results by category?

Yes. This is another powerful feature benefiting from relational technology. You're not limited to presenting results in one long list. For example, if your data has a `state` field, you might want to present the results grouped by state. Categories also may be a hierarchical structure, Yahoo-style.

3.11 Can Texis find related results ("More like this")?

Yes, that is a standard feature. Texis can take any document or text selection and turn it into a search for similar records. This is sometimes called "query by example."

3.12 Can Texis search document "zones" separately?

Elementary! What some people call zones, are in database lingo, fields. With Texis you may query any field separately or in combination with other fields. And queries are not limited to text! If one field (zone) contains a postal code, for example, you could query that with a numeric range such as 90011 through 97000.

3.13 What is the Texis relevance ranking algorithm? Is it tunable?

Texis contains a sophisticated ranking system that may be tuned in various ways. Factors it uses include: closeness of query words to the beginning of a document; order of occurrence of the query words; and proximity (closeness) of query words to each other within a record. These factors may be weighted to change the ranking behavior. As an example of how that might be useful: newspaper articles tend to have the most important material close to the beginning, so in a newspaper search application, you might give that factor more weight.

Chapter 4

Texis for E-commerce

4.1 Does Thunderstone provide a web storefront solution (or auction or classified advertising solution)?

Yes. Generic applications are available from Thunderstone for all of these functions. However, they are not intended as turn-key solutions. Sites of this kind typically each have a unique structure based on the type of products they sell, which would be reflected in a customized database schema and/or some additional application scripting.

4.2 Does Thunderstone offer a web-site personalization solution?

Yes, but note that "personalization" is used to describe many kinds of functionality. At the simplest level, you may store user preferences, so that each user receives only desired information when visiting the site. To accomplish this, Texis can recognize users by login or cookie, and it provides a mechanism for carrying a user's identity from page to page in an encrypted format, so that preferences may be taken into account on every page visited. Personalization sometimes refers to a more elaborate process of capturing user actions (such as purchases or even individual clicks) on an ongoing basis, and using that information to influence what is displayed to the user subsequently. Such logic could be implemented in Texis but is not provided as a finished solution.

4.3 Can Texis search an existing product catalog?

Yes. Whether your catalog exists in another database, or in some other structure, it can easily be imported into a Texis search engine.

Chapter 5

Texis for Web Searching

5.1 Can Texis provide searching of web sites related to one specific industry?

Maybe. If you have a list of the relevant sites, it's easy. But we don't ourselves maintain lists of what are the appropriate sites to spider for any particular topic.

5.2 Can Texis search the web in addition to my local content?

Yes. There are a variety of ways to accomplish that. If you have a specific list of sites on the web you want to search, it's easy. If not, the usual approach is to use a free partner site such as www.master.com for a broad-based web search.

An alternative is to implement a "meta-search" against other search engines. Texis Web Script includes tools for setting that up. Note, however, that other search engines may or may not allow meta-searching.

If you seek to spider the entire web and build a proprietary web-search engine, Texis also is an excellent platform for that; however, such an undertaking takes considerable resources for bandwidth, machinery, administration, etc., and is normally not practical as a "sideline" or by a thinly funded start-up business.

5.3 Can Texis power a bid-for-keyword search service?

Yes, Texis is an excellent platform for such a function. Ideally, this requires a sophisticated trade-off between the relevance ranking and the advertiser's bid for each possible result item. Texis has a facility for tuning and accomplishing such a calculation efficiently.

5.4 Can Texis power an Open Directory Project (dmoz.org) search service?

Yes, Thunderstone provides a generic Open Directory solution that may be customized in many ways.

5.5 Can Texis extract product prices (or addresses or salaries or whatever) from web pages?

Maybe. It depends on whether that information is readily identifiable on each page, either by means of a tag or a predictable structure. If so, then a Texis Web Script may easily accomplish the extraction and save the information as a separate field in the database.

5.6 Can Texis do federated searching?

Yes. A federated search refers to a query that is submitted to two or more search engines or collections, then the results are displayed together or combined. This may also be called a meta-search. Texis Web Script has a comprehensive set of tools for setting up federated searches.

5.7 Can Texis crawl news publisher sites and provide a news search engine?

Yes, Thunderstone provides a generic news crawling solution that may be customized in many ways.

Chapter 6

Texis in Other Applications

6.1 Can Texis search a newswire feed?

Yes. Texis is an excellent platform for news searching. A Texis database can index a news feed in real-time, meaning there is no lag between the time a story arrives and when it is searchable. Texis also provides an efficient mechanism for setting up stored queries for automatic news filtering.

6.2 Can Texis search e-mail or discussion groups?

Yes. Texis is an excellent platform for this. Utilities for importing email as well as usenet (NNTP) data are available from Thunderstone. In the USA, this type of application may be especially important for compliance with laws mandating that significant business data be saved and kept available for auditing. Indexing email archives can also be an important management role. Texis goes far beyond most email archiving solutions that only allow one to search by header such as the `From`, `To` and `Date` fields. Texis will index the full text of both messages and attachments. It will allow sophisticated queries such as a phrase in attachment, between two dates, with results grouped by author.

6.3 Can Texis search graphics or other binary content?

Yes, assuming those files have some descriptive text (metadata) such as photo captions or song titles. Texis does not know how to look at a photo of a giraffe and recognize it as an animal, unless it is labeled as a giraffe!

6.4 Can Texis search scanned documents?

Yes, if they have been "OCR'ed" (converted to text by optical character recognition). Texis does not perform the OCR function. That must be accomplished with document conversion software from another vendor.

6.5 Can Texis classify documents into categories?

Yes. In the simplest case, this depends on how well you can define the categories. If you can specify a query or set of words describing each category, it is quite straightforward. Assigning a category does not need to be a yes-or-no operation; multiple categories may be assigned with a strength (relevance) rating. In many cases that we see, the source or author of a document, or other metadata, plays a significant part in the rules determining category assignments.

In other situations, the rules for assigning categories are not so clear cut. Categories may have been assigned in the past by humans who simply "know one when they see it." For these needs, the Texis Categorizer is available as an add-on module. The Categorizer "learns" to reproduce these decisions from past the category assignments. It assigns the most likely categories, each with a relevance score. New categories may be created by administrators on an ad-hoc basis.

Chapter 7

Software Compatibility Issues

7.1 What other software is required to make use of Texis on the web?

None other that the operating system and web server! Texis provides a suite of tools that together enable a designer to create, deploy, and maintain many web-based applications.

7.2 Which web browsers and web servers is Texis compatible with?

All browsers and most web servers work fine with Texis.

7.3 Can Texis search an Oracle database (or SQL Server or Sybase, etc.)?

Yes! Many of our customers use Texis side-by-side with another database. Using two databases in this manner is straightforward because both obey the SQL standard and thus they may easily exchange data – in real time if necessary.

7.4 Does Texis use XML?

It can, but it is not required. XML was developed principally as a data interchange mechanism, and it is useful for acquiring data from, or providing it to, an affiliate site. On

output, Texis can apply XML markup dynamically even if the data is stored without XML tags. On input, XML data typically would be parsed into fields; but XML mark-up can be preserved in the database if necessary. Texis Web Script contains facilities for full manipulation of XML data. Webinator (a product built on top of Texis) has a SOAP and XML API for searching.

7.5 Does Texis use metadata?

Yes. Metadata is information that describes a document, such as author, source, or subject categories. Texis stores any such data routinely, and can enable searches on this data separately or together with the body text. Texis also make use of metadata for sorting or grouping text search results.

7.6 What platforms does Texis run on?

Texis runs on Linux and Windows Server, and other major OSes depending on demand. For an up-to-date list of supported platforms, see the Download Webinator page on our web site, or contact a Thunderstone sales representative.

7.7 What hardware resources does Texis need?

Resource requirements depend on the database structure, record count, record sizes, query complexity, and of course any other functionality handled at the application (script) layer. As a benchmark, a Texis database containing one million records of typical web page content can serve typical web-search queries at a sustained rate of at least 10 per second on a single-CPU Unix server with 1GB RAM. The biggest single variable affecting performance is usually RAM, which will be used for index caching. Disk space needs are approximately the ASCII text size of the content plus 20 percent for basic full-text indexes. The space needed for indexes could be up to several times more for applications requiring additional or more detailed indexes.

7.8 Does Texis need a dedicated server?

Not necessarily. Texis can share a server with other software such as a web server or another database. It just depends on the resources available in comparison to the combined workload of everything running on the machine.

7.9 Can I use Texis if our site is hosted by a third-party service?

Yes, if you have a dedicated machine at that hosting service. If your site is run on a shared machine, hosting services usually frown on your installing third-party software.

Chapter 8

Application Development Issues

8.1 How do I interface to Texis? Is there an API?

There are a variety of ways to connect an application program to Texis, but the most common is by HTTP. This has the advantage of being simple and high-performance.

A feature-rich, C-callable API is available for special situations, but for most web applications, the HTTP interface is easier to use and almost as fast.

8.2 Does Texis have an ODBC interface?

Yes. ODBC (Open Database Connectivity) is a protocol developed by Microsoft as a generic interface to any relational database. Unfortunately, the specification has considerable "overhead" that tends to make it slow, and it may not be suitable for high transaction rates. HTTP is much faster for those situations. Texis also provides a DBI-DBD interface (Perl module), but performance cautions similar to those of ODBC apply.

8.3 What is the typical implementation effort or time to make use of Texis?

A simple search of web pages, with customized user interface, might take a couple of days. A search of a typical product catalog would typically take a week of scripting effort to get to prototype stage, and another week for refinements. More elaborate applications usually take no more than a month.

8.4 What experience or skills are needed to set up a Texis search engine?

There are two relevant skills. One is familiarity with the SQL language for database application development. The other is either simple programming experience, especially using loops; or html scripting, which may be with a wide range of tools.

8.5 Can I customize the Texis results presentation (look-and-feel)?

Yes, completely! Texis may be used as a "back-end" technology, and imposes no requirements as to user interface. Texis Web Script, a tool for creating web applications, is "neutral" with regard to what HTML mark-up (or other user interface technology such as JavaScript) is used for the user presentation. Webinator supports complete customization of results output via administrator-created XSL stylesheets.

8.6 Does Texis have a graphical user interface (GUI)?

Yes, Webinator and the Search Appliance (products built on top of Texis) have web-based GUIs. Texis itself does not; the set of application development tools provided with Texis use the SQL and HTML programming languages. Texis scripts are created as text files containing special HTML tags for defining database interaction. The HTML look-and-feel may be created with any HTML authoring tool, then combined with the tags defining the database calls.

8.7 Can I interface Texis to a server-side Java (or Perl or VB or ASP etc.) application?

Yes. Any web application program can communicate with Texis via HTTP. Use of Texis Web Script for creating the user interface is optional.

8.8 Can Texis index a combination of web pages and database content?

Yes. Search results may be presented together or separately, as desired.

8.9 Can Texis restrict groups of people from seeing certain docs?

Yes. This capability is very flexible. Entitlements may be set document-by-document, or by groups of documents. Unauthorized users will not see restricted documents in search results and will not be able to retrieve them any other way. Texis recognizes users by login or cookie, and it provides a mechanism for carrying a user's identity from page to page in an encrypted URL, so that entitlements may be taken into account on every page visited.

8.10 Does Thunderstone provide training in the use of Texis?

Yes, although that may not be necessary. Tutorials and tech support often provide sufficient background to become proficient. Training classes, if desired, usually are organized on a custom basis and conducted at the customer site.

8.11 Does Thunderstone install and configure Texis?

Ordinarily, the customer can accomplish this with the possible help of Thunderstone's technical support group. However, consulting services are available for special situations where needed.

Chapter 9

Performance Issues

9.1 How well does Texis scale up? What are the benchmarks?

Texis is by far the highest-performance product in the marketplace providing full-text search within a relational database framework. It powers some of the largest search sites on the internet. Texis provided the search engine at eBay from their earliest days, and scaled up to serve more than 40 million searches a day. Databases of tens of millions of records are not considered large.

9.2 How many documents or records can Texis search?

There is no inherent limit. Texis is routinely used on the most heavily trafficked web sites for searching databases of tens of millions of large records. It has been used with hundreds of millions of records with no significant complications.

9.3 How quickly are Texis text indexes updated?

Instantly! Texis performs standard database record locking, unlocking, and management of contention. It keeps the data consistent and available for all users while records are being inserted, updated, or deleted. No other search engine performs these database-type functions.

9.4 Does Texis do incremental indexing?

Yes. Items added to the database are searchable instantly. Texis takes care of all index updating in background.

9.5 Isn't CGI scripting slow? How can Texis be fast and use CGI?

One of the methods of searching Texis data is through a CGI script. CGI inherently is a very simple, very efficient mechanism. However, it has become associated with Perl and other interpreted scripts that are relatively slow due to the interpreter overhead used at every invocation. Texis scripts are compiled and very fast.

9.6 Is Texis fault-tolerant?

Yes. At the user's option, Texis can run in a fault-tolerant mode. Whenever changes are being made to the database, backup data enabling recovery are preserved in case of power failure or similar disruption.

9.7 Can Texis be used in a distributed/clustered/redundant architecture?

Yes, most of our larger customers have implemented some sort of redundancy strategy. The designs tend to vary based on many particulars including the size of the application, user load, frequency of update, etc.

Chapter 10

Linguistic Issues

10.1 Can Texis search data in languages other than English? Does it handle the "accented" characters of Spanish or French etc.?

Yes, Texis is used in many languages. It probably will work automatically for most European languages. Support for UTF-8 (Unicode) is standard. For other encodings, some configuration settings may be necessary. Accent characters and any other non-English characters will be preserved in the data and become fully searchable, if desired. There are settings to control how searches respond to case, diacritical marks (accents etc.), ligatures, character width etc. For example, searches may be configured to ignore accents, so that users unable to enter accented characters may still find accented-character data, and vice-versa.

10.2 Can Texis index multi-byte languages (Chinese, Japanese, etc.)?

Yes, Texis has been used in these languages. A simple configuration setting tells Texis to index multi-byte patterns (if other than UTF-8). Our customers report satisfactory results with this approach. However, there are some options to improve the accuracy. For example, a specific character in Chinese may sometimes be a word on its own, and other times part of a different word. Chinese readers discern the difference from the context, but there is no indication in the text as to which it is. If you need to index one of these languages, please contact us to discuss these and related issues.

10.3 Does Texis do "stemming"? How about in other languages?

Stemming refers to a process of stripping a word down to its root by removing suffixes or prefixes (such as the "s" on the end of English plurals), and then searching for valid variations of the root (known as morphemes). Texis provides very sophisticated morpheme processing, with default rules that apply to English. Various aspects of morpheme processing may be turned on or off, and the rules customized. A set of morpheme processing rules may be specified for any language. However, we are not linguists and have not defined the stripping and rebuilding rules for most other languages. A user organization typically will wish to customize these rules not only for your language, but for a particular type of data or search style.

10.4 Does Texis have a thesaurus capability?

Yes, a very extensive one. The thesaurus may be customized for any special subject.

10.5 Does Texis search noise words (such as "the" or "is" etc.) ?

Yes or no, as desired. The default behavior is to remove noise words from queries. However, the noise-word list can be customized or turned off, allowing users to search for "the" or any word, if needed.

Chapter 11

Other Technical Issues

11.1 Can Texis search according to geographical locations, such as zip code?

Yes. Texis is unique in its ability to store text records containing geographical locations, and efficiently perform a text search restricted to some distance from a particular point ("swimming pool repair within 10 miles of Columbus, Ohio"). This is accomplished by converting the locations into by longitude and latitude. More details are available from Thunderstone.

11.2 Can Texis index dynamic content such as JSP, ColdFusion, or PHP pages?

Yes. Dynamic content usually implies that data is stored in a database. Texis can easily make any database content searchable, and can provide the search engine behind a Java application.

11.3 Does Texis do agent searching?

Yes. This usually refers to a query submitted by the user and stored on the server; a process notifies the user whenever new data is found. We call this stored query a "profile." Texis contains a very versatile profile processing capability. It can handle a high volume of profiles, matching them in real-time against either incoming data (such as news); or against the results of a web crawler process; or against any other data source.

11.4 Can Texis do a sub-search i.e., only within previous search results?

Yes!

Chapter 12

Business Issues

12.1 How much does Texis cost?

Entry-level Texis prices are in the range of US$8K to $14K one-time. Prices are scaled according to two variables: (a) maximum records per table, and (b) maximum transactions per day (usually corresponding to Texis results page views per day). Leasing also is available. Please contact us for more details.

12.2 How can I get a trial or evaluation copy of Texis?

We provide a free downloadable example application built on Texis. That is Webinator. It includes the source code of the user application layer so that you may examine or modify the SQL `SELECT` statements and other functionality.

12.3 Is there a developer version of Texis?

Every Texis license comes as a bundle with a complete set of developer tools. An entry-level Texis license may be considered the developer version.

12.4 Is there a single-user version of Texis?

Although Texis will run on a PC or workstation, we do not offer single-user pricing. Texis is fundamentally a server product designed for a network. Texis licensees are free to install the software on multiple machines at no additional cost, so it is common for application developers to install copies on their desktop machines.

12.5 Can I get a site license for Texis?

All Texis licenses are site licenses in the sense that the software may be installed on an unlimited number of machines at the customer's site. See above about the price structure. (Other Thunderstone products may be licensed differently.)

12.6 Does an application service provider (ASP) need a separate license for each web site?

No. Applications hosted at one site are bundled under one master license.

12.7 Is there special pricing for educational or nonprofit institutions?

We offer Webinator in part to meet the needs of this market. Webinator provides an actual Texis database with a subset of Texis functions, and the entry-level version is free! Many universities use Webinator for site-search functions.

12.8 Must Texis customers display a Thunderstone emblem on their sites?

No. Texis customers may totally customize the look-and-feel of their applications, and a Thunderstone acknowledgement in the HTML content is not required. However, Webinator users must display a Thunderstone copyright and logo.

12.9 Does Thunderstone have a VAR (value-added-reseller) program?

Yes, we encourage VAR relationships. The key aspect is in the "value added." We expect our VARs to add value by developing applications on top of Texis for third-party clients. Please contact us for more information on becoming a VAR.

12.10 Does Thunderstone have a European representative?

Yes. See the reseller information page at:
`http://www.thunderstone.com/texis/site/pages/Resellers.html` Thunderstone also sells and supports Texis directly worldwide from our USA location. Please contact us regarding your requirements.

12.11 Will Thunderstone develop an application or solution to my specifications?

Maybe. We do have a consulting arm that undertakes such efforts. The requirements need to be well developed before we can bid on a project. A good way to define the requirements is to mock up the entire application in static HTML pages. We are not graphic designers and do not create the look-and-feel. But we can incorporate a design scheme defined by others.

12.12 Can Thunderstone recommend any third-party Texis consultants?

Maybe. There actually is some misunderstanding about what skills Texis consulting entails. The relevant experience is mostly related to standard rdbms application development and web scripting. Very little of it is Texis-specific. Thus, the best consultant may be a local database application developer who can sit down with you to understand your project. Ideally, a consultant will have implemented a web front-end to a large SQL database, such as one involving a million-row table. With that in mind, however, please let us know your needs. We might be able to refer you to a consultant or VAR for a custom solution to a specific problem.

12.13 Is Thunderstone stock publicly traded? What are Thunderstone's annual revenues?

Thunderstone is privately held and does not disclose internal financial information. But we can tell you that we are a 30-year-old company that is conservatively managed and has never required venture capital. We do not subscribe to the philosophy of trying to build market share by spending millions of dollars more on promotion than comes in as revenue. We have

grown based on re-invested profits, and we stay away from "bet-your-company" strategies!

12.14 What fallback do I have if Thunderstone ever goes out of business or discontinues supporting Texis?

Texis is a mature and stable product that has been on the market for 17 years. At most customer sites, Texis runs for many months without restarting. However, we recognize that many customers are wary of becoming overly reliant on any one software vendor. The ultimate fallback, should a customer ever need to discontinue using Texis, lies precisely in Thunderstone's adherence to standards. Texis data may easily be transferred into any competitive SQL database. Likewise, Texis scripts consist essentially of standard HTML with encapsulated SQL calls that could, with some effort but not an unreasonable amount, be converted to a competitive scripting environment. Of course, other databases at this time do not come close to Texis's performance in text-intensive applications. However, if that situation changes in the future, the migration path to an alternate solution would be straightforward.

12.15 Does Thunderstone provide a hosted solution?

Yes, Thunderstone Data Services provides hosting services to Texis customers. Please contact us for further details.