



SMASER

TRIP Overview

White Paper

Abstract:

The TRIP family of products are associated with providing high performance storage and retrieval of a full range of information and data types. These can include structured tabular data, structured text, free text, and virtually any type of locally or remotely stored objects. This overview is designed to provide a description of the basic architecture of the TRIP family of products, their major functions, and the various configurations available.



End User License Agreement

All rights to this software, its documentation and logotypes of the TRIP product family and software (altogether "Software") supplied by Smaser AG (Smaser) are exclusively owned by Smaser.

The transfer of this Software, solutions or parts thereof requires the prior written agreement of Smaser. Furthermore, the customer has the right to use licensed Software and / or process solutions supplied by Smaser to the extent specified in his contract with Smaser.

The free-to-use non-commercial version doesn't require a prior written agreement with Smaser but such customers, organizations and/or third parties agree by using the software and / or solution of Smaser to be strongly obliged to keep all rights to this software, documentation and logotypes of the TRIP product family absolutely un infringed and protected.

Table of Contents

| | |
|----------------------------|-----------|
| INTRODUCTION..... | 4 |
| TRIPSYSTEM | 4 |
| DATABASE DESIGN..... | 4 |
| USER MANAGEMENT | 5 |
| SECURITY | 5 |
| DATA ENTRY..... | 6 |
| SEARCH AND RETRIEVAL | 7 |
| CCL..... | 7 |
| SQL..... | 8 |
| XPath | 8 |
| REPORTING | 8 |
| TRIPMANAGER | 9 |
| TRIP SDK | 10 |
| SUMMARY | 11 |

Introduction

The TRIP family of products is built around a high-performance text database engine known as TRIPsystem, which is supported on a several operating systems. Developers access the engine via TRIP's SDK, which encompasses a wide variety of access protocols, including stand-alone, client/server, 3 tiers and web, and a wide variety of host languages, including C/C++, Java and .NET. The features of TRIPsystem are also exposed through an XML-based interface.

Enterprise administration of heterogeneous TRIP server installations is provided by TRIPmanager, a Microsoft Management Console snap-in. End users typically interact with TRIP via a custom-written application or via an out-of-the-box application for a specific purpose.

TRIPsystem

At the heart of the TRIP family of products is TRIPsystem, a database engine with a fully open architecture which provides all of the functionality needed for building, maintaining and managing applications, storing and retrieving information, and managing access and security. The engine uses a dual indexing strategy based on a positional inverted index combined with a three-level n-gram index. This combination provides maximum performance and flexibility for the handling of literal or truncated Boolean and positional searches as well as non-Boolean algorithms such as fuzzy logic, approximate matching, concept-based searching and query-by-example.

TRIPsystem is optimized for handling large amounts of data, as well as complex heterogeneous combinations of data types within applications. Databases designed using TRIPsystem have essentially no boundaries with regard to the amount of data per field or record, number of fields or records per database, number of databases per system, or the size of any database (other than the limits of the computer system on which they are installed).

TRIPsystem uses a wide band hash-coding algorithm that all but eliminates "collisions" in indexed terms. This gives all TRIP products an advantage in both accuracy and precision when called on to perform complex search strategies, without sacrificing overall performance. Additionally, TRIPsystem is Unicode enabled, thereby providing search and retrieval support for nearly all language groups.

TRIPsystem has the unique capability of permitting simultaneous access to multiple databases without the need for common key fields or links as is typically required by relational model systems. It thus provides the maximum flexibility available in the industry today with regard to the size and/or scope of applications it can support. Further, TRIPsystem's architecture makes it possible to produce completely interoperable applications with little advanced planning.

TRIPsystem is supported on Microsoft Windows, Oracle SPARC Solaris, IBM AIX and Linux. The result is compatibility across the full range of platforms right down to binary compatibility of the database files and the data contained therein.

The following discussions highlight TRIPsystem's major functions.

Database Design

There are two major database types that are supported by TRIPsystem: XML and flat. In the former case, the database design is optimized for path-specific searching. In both cases, the content of the database is defined by the record structure, consisting of an

optional master/detail layout and as many structured or unstructured fields as are necessary to fulfil the application's requirements.

TRIPsystem supports data storage for a variety of data types, including unstructured text, short structured text, binary objects, and the usual array of date, time and numeric formats.

In addition, TRIPsystem supports language classes that determine how the system will interpret characters for various language types. For example, the English character class interprets the characters "o" and "ö" as the same character, while the German character class treats them as being different. Similarly, the Swedish character class not only treats those two characters separately, it appropriately respects the sort order used by the Swedish language, whereby the "ö" character is sorted after "z". TRIPsystem also allows the DBA to extend the character class with additional searchable characters (e.g., "+", "-", "/", etc.) that TRIPsystem does not normally index.

Data validation within a TRIP database can be performed using two mechanisms: internal and external. Internal validation takes the form of either cross-checking with another database (e.g. using a lookup table to validate entries) or using a regular expression for format checking. External validation takes the form of a declared 3rd party function (written in a host language such as C) that can perform whatever type of validation is required and either pass, fail or modify the provided data.

XML data is supported automatically via a simple switch in the database design that informs the kernel of the likelihood of the presence of XML data. With this switch active, and when using the optional TRIPxml data import facility, the database becomes fully XML aware and is available for path-specific searching using element and attribute names from the original XML documents.

User Management

TRIPsystem has powerful facilities for managing users and user accounts. There are four basic user types: System, File Manager, User Manager and End User. "System" is the user type assigned to the system manager, and is the only user that can create File Managers or User Managers, and declare an application "public" (open to all users).

File Managers can create databases and are therefore typically application developers. At most installations there are many File Managers, since end users frequently build their own applications. File Managers are also responsible for granting access rights to other users or groups.

User Managers have the right to create users and groups and to assign users to various groups. TRIPsystem separates user management tasks from access and application control to ensure that maximum system management flexibility is maintained without compromising available security features. In many cases a user may be assigned both File Manager and User Manager rights. TRIPsystem permits user and security management to be tailored to the needs of the organization. In this way, installations can realize the greatest benefit from the system's power and versatility.

Security

Compatible with operating system security features, the security functions begin with the structure of user types as defined in the previous section and extend through the search system and access control to databases and data.

As previously stated, there is a separation between those who can create an application and those who can create users. Once users and/or groups exist, a File Manager can assign rights to any database created by that File Manager. Rights are granted from two perspectives, field level access rights and content-based access rights.

Field level rights determine the extent to which an assigned user or group has read or write access to a given field within a database. Users with read-only access can search and view data in assigned fields but cannot perform data entry or modify data contained within those fields. If read access has been denied, all access to those fields, including reference to them, is restricted. Even a request to list the fields contained in the database will not show references to read-restricted fields.

Content-based access rights can be defined for a given user or group to a given database. These rights restrict that user's view of the data within the database to only those records that pass the content-based restriction (a query on the database).

These combinations of security even make it possible to permit a user to enter data and then be restricted from accessing it once it has been stored.

Data Entry

TRIPsystem supports both interactive and batch data entry and modification. Interactive data entry/modification is accomplished through use of the SDK, and is typically enabled via a custom-written application or one of the out-of-the-box applications.

Batch loading of databases is accomplished using either XML (via the optional TRIPxml add-on product), or via a proprietary tagged-text format called TFORM.

As is always the case, strict adherence to TRIPsystem security applies regardless of the method employed.

Search and Retrieval

Searching a TRIP database is accomplished using either CCL (a derivative of the Common Command Language, NISO std. Z39.58), SQL using the optional add-on product TRIPsql, or XPath (on XML data) using the optional add-on product TRIPxml.

CCL

Standards Based

The Common Command Language was popularized by the WAIS distributed searching protocol (std. Z39.50) and consists of a small number of highly intuitive commands that are suitable either for developer's or end user's use.

The BASE Command

The BASE command opens one or more databases for searching. The databases must have compatible character classes, but need not have any specific key fields or links in common. Therefore, highly diversified data collections can be searched with ease. Up to 250 databases can be opened simultaneously, allowing for extremely large data populations.

The FIND Command

The FIND command does just that. It finds a word, phrase or other data item in a database. For example, a user can simply type "FIND GENERAL MOTORS" to perform a search in all open databases for occurrences of 'General Motors'. TRIPsystem is case insensitive, and no stop words or special constructs are required for phrase-based searches. FIND commands can apply to every indexed field in the database, or be restricted to specific fields (or groups of fields). Since TRIPsystem uses a positional inverted index combined with an n-gram index, all searches are direct indexed retrievals, whether on full words, or using front, back, middle or mixed wildcarding. It is TRIPsystem's dual indexing feature that results in the product's remarkable overall performance, even for complex queries. The FIND command supports all forms of positional searching, including within n words, within a record, entity (a joined combination within a master/detail layout), field, table, subfield, paragraph, sentence, even within the same word. FIND also supports all standard Boolean operators as well as providing several forms of non-Boolean algorithms that combine a concept-based search operation with a ranking algorithm that orders the records found according to their relevance to the original search.

Approximate Matching

TRIPsystem also provides a user tuneable approximate term matching algorithm. Using this feature a user can easily adjust the degree of matching desired, from a tightly controlled match of terms or term fragments to a loosely controlled fuzzy match in which found terms might have as few as two characters in common with the search term. Because the fuzzy matching algorithms are based on the n-gram index, it is easy to understand what they are doing and to maintain a high degree of control over complex searches. This is particularly useful when searching lower quality data, such as might result from OCR (optical character recognition) of scanned images.

Indirect Search

TRIPsystem provides very powerful indirect searching features. Users can refer to previous searches as part of any standard search statement. They can also create complex mapped search strategies where the entered terms are used to search a second database (or group of databases), from which the contents of specified fields in that database are used as the search terms in the currently open database (or group of databases). Highly complex relationships can be defined and used for searching by expert

users, and they may be associated via stored procedures with specific users or user groups so that even novices can benefit from the full power of the system.

Thesaurus Support

TRIPsystem provides a comprehensive thesaurus facility. Its multidimensional thesaurus conforms to ANSI standards as well as providing added functionality for applications requiring a full network-based structure. The thesaurus can be used for classical hierarchical thesaurus applications, synonym lists, and topic-oriented strategies, as well as for more complex applications such as table of contents structures, hierarchical lists (e.g. organization charts, SGML document structures, etc.) and even complex network diagrams (such as tracking computer or phone system cabling networks). TRIP's thesaurus is fully integrated with CCL and enables users to mix direct, indirect and thesaurus search strategies in the same commands if desired.

DISPLAY: Term Lookup (for facets, etc)

The DISPLAY command provides a user with the ability to see a pick list of terms related to the search criteria specified. The syntax is generally the same as for the FIND command, including the use of wildcarding and approximate term matching features. Once a list is displayed, a user need only select one or more terms to create a search statement for execution.

SQL

Many of the features of TRIPsystem that are exposed through CCL are also available using SQL, via the optional TRIPsql add-on product. This product supports access to TRIP via either ODBC or JDBC, depending on the host language environment.

XPath

XML databases in TRIP are best queried using XPath. Search expressions in XPath are executed against all XML documents in the currently open XML database at the same time. Retrieval of XML fragments from individual XML documents or from several XML documents in a search result is also supported. Using XPath queries with TRIPxml is best done via the SDK products TRIPjxp and TRIPnxp.

Reporting

Depending on the search infrastructure being used (either CCL or SQL), the reporting mechanism used may be different. Typically a SQL application will use an off-the-shelf report writer (e.g. Crystal Reports), whereas a CCL application might well follow the same strategy, but more likely will use TRIP's own report writing capabilities to output data to an HTML (for example) device.

TRIP's report writer, which is bundled as part of the TRIPsystem, supports highly flexible formatting of information from TRIP databases and via the TRIPmanager administration console, can be wizard-driven to a high degree so that the production of new reports for HTML, XML or textual output is a simple point-and-click exercise.

TRIPmanager

To manage an installation of one or more TRIP servers, potentially running on one or more different operating systems, administrators use the TRIPmanager product. This Microsoft Management Console snap-in component supports distributed management capabilities for any number of TRIP servers, operating on any supported operating environment. Administrators use TRIPmanager to design new databases, to manage their user population, to create stored procedures and to generate TRIPsystem reports.

All activities within TRIPmanager are either simply point-and-click exercises, or are automated via wizards so as to make the work involved as simple as possible.

TRIP SDK

Building on the TRIPsystem, the TRIP SDK provides a wide variety of means for developers to create novel applications:

- C/C++ development using TRIPapi. This is the lowest level interaction mechanism provided, and whether used directly, or via a client-side proxy to a TCP/IP client/server connection, allows developers access to the fundamental data structures and APIs of the TRIPsystem. All other access mechanisms in the TRIPsdk are built upon the TRIPapi foundation.
- Development in the Microsoft .NET environment (e.g. using C#, VB.NET, etc) using TRIPnpx, a class library built on top of a low bandwidth-consuming network protocol called TRIPxpi.
- Java development using TRIPjxp, a library of Java classes that provides the same object model as TRIPnpx for the Java language environment.
- HTML templating using TRIPhighway, a cost-effective way of exposing TRIP databases over the web in read-only search-and-retrieval applications.

Another component of the TRIP SDK that enables developers to take advantage of rich document types, both for viewing and indexing, is TRIPcof. This component provides support for the most common file formats (including PDF, Microsoft Office, OpenOffice, etc.).

Also available, but considered legacy and end-of-life as per the release of TRIP 6, are tools for:

- Microsoft component-based development (e.g. VB, ASP) using TRIPcom, an in-process COM component that provides a number of objects that encapsulates TRIP concepts in an easy-to-understand model.
- Java development using TRIPjtk, a library of Java classes that provides the same object model as TRIPcom for the Java language environment.

Summary

The TRIP family of products provides all of the tools necessary for application development across a wide range of needs, from systems integrators to end-users. Additionally, the company's distributors and agents offer a full range of consulting services to help organizations design and build comprehensive document management and information systems customized to their needs, yet flexible enough to adapt to and grow with the ever-changing environments in which they operate. The TRIP family of products runs on a wide variety of platforms, offering power and performance for installations large and small, from host-based systems to client/server networks, from networked office systems to stand-alone workstation users.