TRIP Connectivity Framework v1.6-0

Change History

2024-01-17

Web: https://smaser.ag

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 uninfringed and protected.

Table of Contents

About This Document

This document contains the change history for TRIPcof.

Known issues

[TRIP-5453] Moved directories not always detected

The 'fsbot' file system connector does not always notice or completely handle subdirectories that are moved. This can be a move within the monitored tree, a move to a non-monitored area, or thee move of a non-monitored directory into a monitored one.

This can in some cases also result in files being dropped from the TRIP index. This will only result in them not being searchable; they will still be present on the file system.

As a workaround, run the cfwimport program in a non-active mode at regular intervals, using the '-new' option to detect files that may have been missed or dropped due to this issue.

[TRIP-5452] Connector services not retained after upgrade

Connector Windows services services hosted by the cfwimport program that have been created by the admin are not retained after upgrading TRIPcof.

The current workaround is to recreate the services manually after upgrade.

Version 1.6-0

News in v1.6-0

[TRIP-5886] Required Linux is now RHEL9

[TRIP-5866] Java 17 compatibility

[TRIP-5853] Apache Tika dependency upgraded to 2.9.1

Version 1.5-1

News in v1.5-1

[TRIP-5771] Apache Tika dependency upgraded to 2.7.0

Corrected in v1.5-1

[TRIP-5661] PDF-to-HTML conversion leaves temporary files

The ICEpdf file filter adapter which handles conversion of PDF documents to HTML

would emit intermediary data to files under the system standard tmp directory. Those files were in addition not removed after completed conversion. This resulted in an excess of potentially large files in the system's tmp directory. This posed a small security risk as well as risking filling up the file system after prolonged use.

The temporary files are now removed after conversion is completed. A new property "EnableFileCache" for the ICEPdf adapter can now also be set in the icepdf.conf file. This property, if set to "False", will disable this caching altogether. This increases system resource use somewhat but reduces the risk of sensetive data being emitted to the file system, albeit for a short duration.

Version 1.5-0

News in v1.5-0

[TRIP-5675] ICEpdf dependency upgraded to 6.2.5

The version of the ICEpdf library used for conversion of PDF files to HTML has been upgraded to version 6.2.5 along with its dependencies on BouncyCastle (to version 1.71) and Apache Batik (to version 1.14).

[TRIP-5618] Apache Tika dependency upgraded to 2.4.0

The version of Apache Tika used by TRIPcof as its main text extraction provider has been upgraded to version 2.4.0.

[TRIP-5599] Java 11 compatibility

TRIPcof can now be used with Java 11. The older Java version 8 remains supported.

Corrected in v1.5-0

[TRIP-5702] RHEL fapolicyd blocks TRIP binaries

A Linux system where the fapolicyd service is in use will normally behave so that any shared object libraries not located in whitelisted directories will not be loaded by the operating system's dynamic loader. In addition, fapolicyd will normally also block access to binaries explicitly not whitelisted. This resulted in "operation not permitted" or "file not found" errors when trying to use TRIPcof.

The installer script will now update the fapolicyd database by whitelisting the TRIPcof lib and bin directories. Similarly, the uninstallation script will remove the TRIPsystem lib and bin directories from the whitelist. The fapolicyd adjustment is managed by the new bin/chfapolicy_cof.sh script.

Corrected in version 1.5-0:1

[TRIP-5699] Old text remains if file contents has been cleared

Previously extracted text from connectors (e.g. the fsbot file system connector) would

remain in TRIP even after an update to the item that specified that no text existed in the new version of the item.

[TRIP-5684] Text extraction incorrect from longer documents

Text extraction from longer documents could become slightly corrupted, causing some paragraphs to become garbled with text fragments from previous sections in the document.

[TRIP-5557] File system connector may crash on Linux

When using the file system connector on Linux to perform active monitoring of changes in the file system, there was a high risk that the detection of a change would not have any effect or cause the connector process to crash.

Version 1.4-1

News in v1.4-1

[TRIP-5528] Memory overhead improvements for Apache Tika

The integration to Apache Tika has been overhauled to utilize less memory when large documents are processed. While the default Java max heap size has been increased to 1G to allow for greater operational headroom, the processing is now stream instead of buffer based, which results in lower overall memory usage and somewhat improved response times.

Corrected in version 1.4-1:1

Corrected in v1.4-1

[TRIP-5571] Invalid file permissions on Linux/UNIX

The TRIPcof installation script would not apply correct file and directory permissions for some of the created files and directories if the root user's umask was set to deny other users access (e.g. value 0027). If the TRIPcof software would then be run as a non-root user (e.g. via tripnetd/tbserver), operations would fail with error messages such as "Could not resolve TRIPcof installation directory".

Corrected in version 1.4-1:2

[TRIP-5537] TRIPcof ASE may not find location of worker programs

TRIPcof ASE may not find location of worker programs. This can manifest in one of two ways:

- A crash, with an "Unexpected state" exception when text extraction or HTML conversion is requested from TRIPjxp or TRIPnxp.
- An error message "TRIPcof worker process exited prematurely with code #1". Examining the fifiase log file in this circumstance shows an incomplete path to tvextract

(e.g. Extractor is to be executed; "/bin/tvextract")

[TRIP-5519] Connectors fail to load in Docker

When running TRIPcof under Docker, connectors (e.g. the fsbot file system connector) would always fail to load.

Corrected in version 1.4-0:6

[TRIP-5514] File system connector file type check on UNIX

The Excludes and Includes settings in the fsbot.conf file for the file system connector were not honored when running the file system connector under Linux and Solaris. All files in the designated folder would be processed.

Corrected in version 1.4-0:6

[TRIP-5512] Text extraction fallback attempt always fails

If multiple adapters are available that can extract text from a particular file format and the first adapter fails, all subsequent adapters would also fail.

Corrected in version 1.4-0:5

[TRIP-5509] Text extraction error message not propagated

If the text extraction failed, the error message was not propagated from the tvextract program via the ASE to the caller (e.g. TRIPnxp or TRIPjxp). Only an anonymous "text extraction process failed" error message was sent back.

Corrected in version 1.4-0:5

[TRIP-5508] Text extractor out of memory error not handled

If the Apache Tika text extractor ran out of memory, the associated Java OutOfMemoryError exception was not properly handled. This caused the text extraction to fail without sending any response message, resulting in an extraction failure due to timeout.

Corrected in version 1.4-0:5

[TRIP-5494] Config not properly initialized under Docker

The TRIPsystem tdbs.conf configuration file was not properly updated with regard to TRIPcof during Docker container startup. Unless configuration was manually updated, this would result in that the the TRIPcof file filter functionality could not be invoked.

Corrected in version 1.4-0:4

[TRIP-5493] Default config may cause text extraction failure

If the TRIPcof configuration files were not located under the installation directory, the

TRIPcof ASE routine would in some circumstances cause the ASE to fail to launch the text extractor child process.

Corrected in version 1.4-0:4

[TRIP-5483] Apache Tika based text extractor may fail to start

The tikamond background process that is used to start and monitor the execution of the Apache Tika based text extraction service would fail to start if the (server-side) process was launched as user without write access to the TRIPcof installation directory.

Corrected in version 1.4-0:3

[TRIP-5482] Risk for crash when storing item from connector

There was a risk for a crash in the TRIPcof registry library (libcfwreg.so on Linux and Solaris and cfwreg.dll on Windows) when storing to TRIP data obtained via a TRIPcof import connector.

Corrected in version 1.4-0:3

[TRIP-5481] Error reading file permissions on Linux and Solaris

The file system connector failed to read file permission information on Linux and Solaris. An error message "Error reading file permissions" was emitted, along with an arbitrary status message (e.g. "no such file or directory", "success", etc).

This would result in that no ACL information about processed files would be added to the TRIP record for the file.

Corrected in version 1.4-0:3

[TRIP-5474] Modified check in cfwimport not working on Linux, Solaris

Running the cfwimport program with option "-modified" to detect modified files would not work on Linux and Solaris. This was due to the file time stamp not being properly handled by TRIPcof.

Corrected in version 1.4-0:2

[TRIP-5472] Extraction from empty file caused hanging

Using the default tikaserver adapter to extract text from a file with zero length would cause it to hang indefinitely with the associated Java process consuming a significant amount of CPU.

Corrected in version 1.4-0:2

[TRIP-5464] Bad encoding of file names imported via the fsbot connector

National characters (e.g. umlauts) were not properly converted to the TRIP session character set by the file system connector, resulting in any umlaut characters the I_ID

field value in the target connector database to be represented as a question mark or not appear at all.

Running the corrected version of the fsbot connector may temporarily result in duplicate entries for files whose names have national characters in them that previously have been imported into TRIP. Resolve this matter by running the cfwimport program on the relevant fsbot data source in delete mode (using the "-delete" option).

Corrected in version 1.4-0:2

[TRIP-5463] Incorrect detection of multiple changes to single file

The file system connector running in active monitoring mode under Windows (as service) would sometimes get multiple file change notifications for a single file, even though the user only performs a single change. E.g., when copying a file into a monitored directory. This would result in the same file being processed multiple times unnecessarily, effectively slowing things down.

Corrected in version 1.4-0:1

[TRIP-5462] Many simulataneous changes hung the fsbot connector

When running the file system connector (fsbot) in active monitoring mode (as service) under Windows, there was a risk for it to hang with high CPU load.

Corrected in version 1.4-0:1

[TRIP-5461] Some file patterns not matched by the fsbot connector

Certain patterns for files to include or exclude as specified in the configuration files for data sources for the fsbot connector could not be successfully handled, even if the pattern was supposed to match a file. For such pattern matching errors, the fsbot connector would (for include patterns) ignore the file, or (for exclude patterns) process the file anyway.

Corrected in version 1.4-0:1

[TRIP-5459] Adapter for Apache Tika not starting on Solaris

The file filter adapter for Apache Tika would not start properly on Solaris. The log file would contain an entry stating that the tikamond program could not be accessed, specifying an incorrect path to it. If started manually, the tikamond program would work and text extraction could be done, but when requested to shut itself down (by running "tikamond -k"), it could not do so, claiming it was not running.

Corrected in version 1.4-0:1

[TRIP-5457] JavaVM not configured for connectors

The cfw.conf configuration file contains a property JavaVM that should refer to the fully qualified path to the Java virtual machine to use with Java based connectors. This propertly was not properly set during install, resulting in longer load times for Java

based connectors.

Corrected in version 1.4-0:1

Version 1.4-0

News in v1.4-0

[TRIP-5437] API package/namespace names changed

The package names in the TRIPcof Java APIs for file filters and connectors have been altered. The prefix "de.infinitservices" has been changed to "ag.smaser".

The namespace names in the TRIPcof .NET APIs for file filters and connectors have been altered. The prefix "InfinITServices" has been changed to "Smaser".

[TRIP-5428] Relocatable configuration and log directories

To better support deployment under Docker, the configuration and log files can now be placed on a different location. Refer to the installation guide and the operating manual for more details.

[TRIP-5423] Update third-party components

The following third party components have been updated:

- libcurl: updated to version 7.75.0
- Apache Tika: updated to version 1.26

Corrected in v1.4-0

[TRIP-5440] Text extraction fails for small OOXML documents

Extracting text from OOXML documents (docx, pptx, xlsx) could fail with a "truncated zip" error message. This affected documents smaller than 20MB.

[TRIP-5439] Risk for indefinite hanging in TRIPcof ASE

When the TRIPcof ASE launches a worker process (for text extraction or HTML conversion) it waits for the worker to establish a connection back to the ASE for inter-process communication purposes. If the worker fails and exits before it has a chance to connect to the ASE, the ASE would hang indefinitely.

[TRIP-5438] Config initialization may fail

TRIPcof allows for the configuration file directory to be overridden by defining the COF_CONFIG_DIR logical name in the tdbs.conf file. If this value is incorrectly specified, no configuration would be loaded, resulting in various negative side effects including hangings.

Version 1.3-0

Corrected in v1.3-0

[TRIP-5344] Zero index interval causes crash in active mode

The cfwimport program would crash when running in active mode if the IndexInterval property (in cfw.conf) was set to zero.

Version 1.2-1

News in v1.2-1

Background database indexing

For import connectors run under cfwimport in active monitoring mode, the imported data will now be indexed at an interval determined by the IndexInterval property that can be set in the cfw.conf configuration file.

Corrected in v1.2-1

Directory deletion not detected in active mode

The deletion of a subdirectory in a directory covered by the 'fsbot' file system connector running in active monitoring mode was not detected. This could cause previously indexed files to remain searchable even though they no longer were present on the file system.

Version 1.2-0

News in v1.2-0

Server-side import connector framework

A connector framework that supports server-side import connectors is now available. A ready-to-use connector "fsbot" for file system indexing is included. Custom connectors can be implemented using the included API for Microsoft.NET 4.

Corrected in v1.2-0

Invalid national characters in document properties

Values of extracted document properties that contained national characters would in some cases be incorrectly encoded, resulting in such words having unreadable characters when stored in TRIP.

Invalid national characters in extracted text

Extracted text from documents containing national characters would in some cases be incorrectly encoded, resulting in such words having unreadable characters when stored in TRIP.

Cfwimport Option -logtofile not working

The -logtofile option for the cfwimport program would result in an empty log file and no output to the console. The import operation still performed normally, though.

Tika server startup race

There was a race when starting the Tika server for text extraction, such that the first text extraction request could be sent before the Tika server was properly up an running. This would result in text extraction attempts failing until the Tika server startup had completed.

Property names from IFilter

The document property names returned from the IFilter adapter during text extraction did not return properly canonicalized names. This could result in properties not being properly named, or given the wrong names.