Antonio Montanana
Software Engineer
October 2014

Product Implementation Training (PIT)

Content Platform Engine 5.2.1

Pluggable Video Transcription Framework

Introduction

- ## Course Overview

  Introduction to the Pluggable Transcription Framework configuration and capabilities

- ## Target Audience

  IBM employees working with partners to integrate a Video Transcription service with P8

- ## Suggested Prerequisites

  Basic knowledge of the P8 Content Platform Engine (CPE), P8 Content Based Retrieval (CBR) and Administrative Console for Content Platform Engine (ACCE)

- ## Version Release Date

  October 2014

Course Objectives

After this course you will be able to:

- Describe the components of the Transcription Framework and how they interact to transcribe video documents and index the resulting transcript to provide full-text search capabilities of the video document.
- Have a basic understanding of the installation/configuration of the framework components.
- Understand how Transcription Queue Sweeps and Queue Sweep Jobs are used in the video transcription process.
- Know how to locate the technical document that provides in-depth configuration details.

Course Roadmap

- ➔ Transcription Framework Overview
- Configuration
- Transcription Queue Sweep
- Demo: Configuration and Transcript Generation
- Best Practices
- Course Summary

Transcription Framework Overview

- The Transcription Framework is part of a P8 CPE technology initiative feature to provide automatic transcription of video documents. This feature, in conjunction with a vendor supplied transcription handler plug-in, provides the ability to transcribe video document content to Timed Text Markup Language (TTML) which is added to a document as an transcription annotation. Optionally, text from the transcription annotation can be extracted and indexed to allow full-text search of the source video document.

- The video transcription feature introduces some new components in CPE v5.2.1 and also leverages existing P8 components:

  Rich Media Transcription (Rms) Extensions: A new P8 CPE Add-on that installs the class definitions, properties and instance data required for video transcription onto an object store.

  Transcription Request Sweep: The existing CPE custom queue sweep feature has been extended in CPE v5.2.1 to provide a Transcription Request Sweep capability by way of a Transcription Request Handler. The request handler transcribes newly checked-in video documents using an event-based subscription and transcribes existing video documents via transcription sweep jobs.

Transcription Framework Overview (cont.)

Transcription Handler: Is a vendor-supplied code module that plugs into the Content Conversion Subsystem to provide the actual conversion from video to TTML.

Transcript Annotation: A new annotation class (RmsTranscriptAnnotation) is installed by the Rms Add-on. After a video is transcribed to TTML, the Transcription Sweep Handler adds a new transcription annotation containing the TTML to the (source) document. The content of the annotation can be indexed for full-text search by using the Transcription Annotation Preprocessor to extract the text.

Transcription Annotation Preprocessor: New in P8 CPE is the capability to extend text extraction for full-text search. The Transcription framework provides a built-in preprocessor for extracting text from TTML transcriptions.

Transcription Framework Overview (cont.)

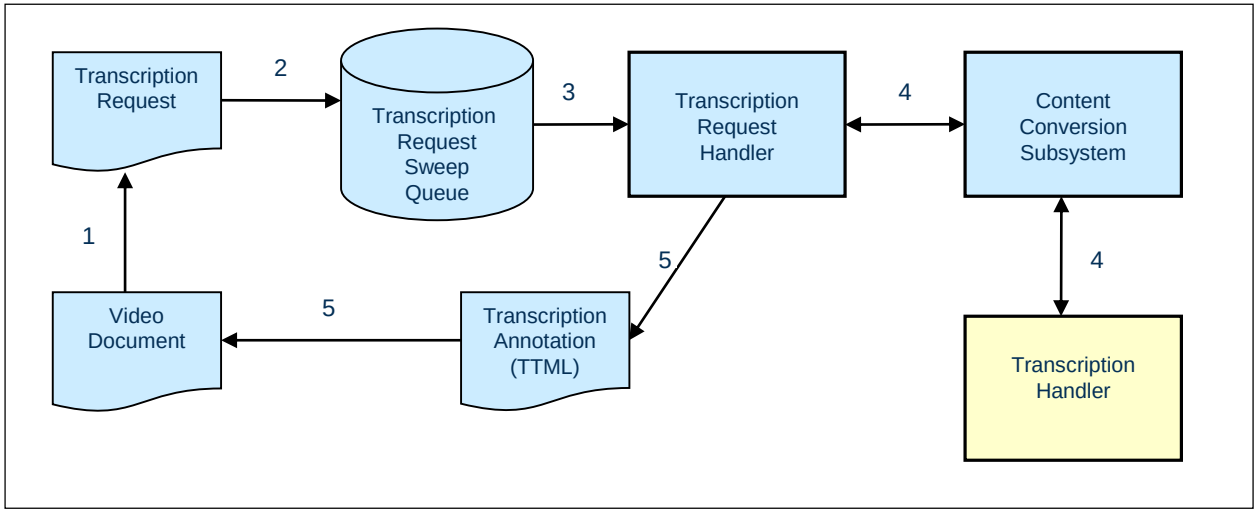

figure 1. Blue: Content Platform Engine, Yellow: Vendor supplied code

1. A transcription request is generated for a new or existing video document by way of a subscription or transcription sweep job. The request could also be generated “manually” by an outside application.

2. The transcription request is added to Transcription Request Sweep queue.

3. Batches of transcription requests are picked up the Transcription Request Handler, the video content of the document referenced by the request is passed to the Content Conversion Subsystem.

4. The Content Conversion Subsystem examines the registered content conversion handlers to choose the appropriate handler for the transcription. When found, the video content is passed on to the Transcription Handler to perform the video-to-TTML conversion and the TTML is returned back to the request handler.

5. The Transcription Request Handler creates a transcript annotation on the document containing the TTML.

Transcription Framework Overview (cont.)

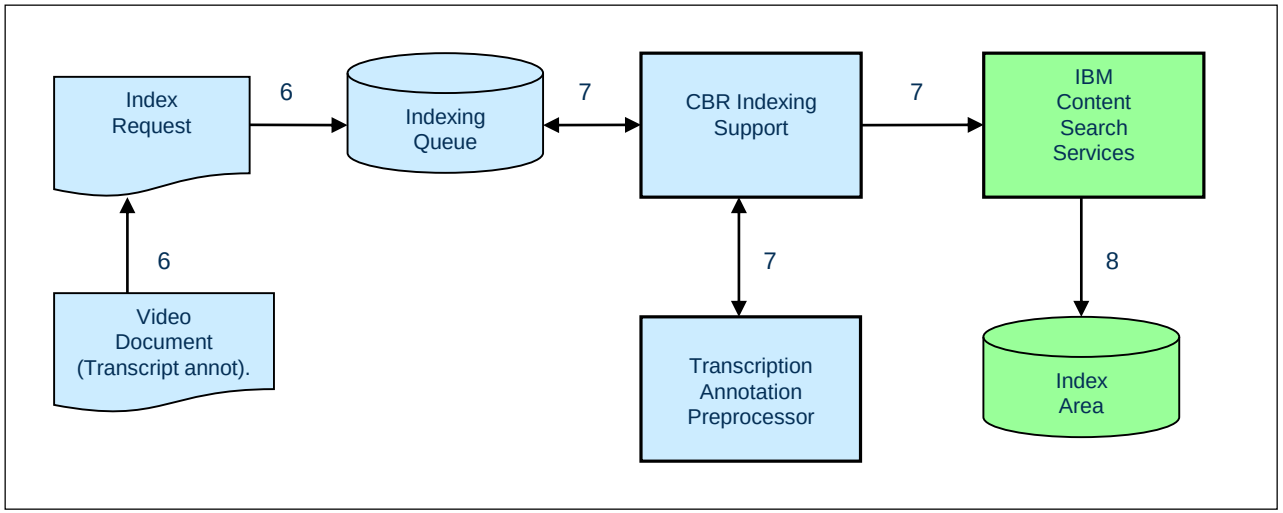

figure 2. Blue: Content Platform Engine, Green: Content Search Services

6. Creation of a transcript annotation (TTML) on the video document triggers an index request to be created and placed in the indexing queue.

7. CBR indexing code pulls the request from the queue, invokes the Transcription Annotation Preprocessor to extract the text from the TTML, the text is then sent to IBM Content Search Services for indexing.

8. IBM Content Search Services indexes the text from the annotation providing a full-text search capability for the video document.

Course Roadmap

- Transcription Framework Overview
- ➔ Configuration
- Transcription Queue Sweep
- Demo: Configuration and Transcript Generation
- Best Practices
- Course Summary

Configuration

- There are several steps required, using the Administration Console for Content Engine (ACCE), to enable the transcriptions on an object store. Details of each of the following steps can be found in the *Transcription with FileNet Content Manager: Configuring Transcription and Building Transcription Handlers* technical document referenced in the Product Help/Documentation/Resources section at the end of this presentation. In general the following is required:

  Rich Media Transcription Extensions Add-on: This Add-on must be installed first, it adds the extensions (class definitions, properties, instance data) required for generating transcriptions to new or existing object stores.

  Transcription Handler: Must be obtained from a vendor and installed. Handler installation and configuration details to be provided by the vendor.

  Transcription Document Class: Creation of a separate transcription document class for video transcription is an optional but recommended as is using file-based storage for video documents (see best practices).

Configuration (cont.)

Transcription Subscription: Transcriptions requests can be automatically created when a new video document is checked in by using the Transcription Request Event Action installed by the Rms Add-on. A Transcription Subscription (subclass of Subscription) must be added to the Transcription Document class with both a Checkin Event trigger and a filter expression defined that limits creating transcriptions for video document content, for example the filter:

VersionStatus = 1 AND (ContentElementsPresent INTERSECTS ('video/mp4'))

will limit create the request for released document with “video/mp4” content elements.

Transcription Annotation Preprocessor: To allow a video document to be searchable it's transcription text must be indexed. This is done by first CBR enabling the Transcription Document class and then associating that class with the Transcription Annotation Preprocessor (installed earlier by the Rms Add-on).

Configuration (cont.)

- Additional configuration steps may be required:

Temporary File Path: The Transcription Handler may require use of temporary files, a directory for these files can be configured for the domain. This configuration will depend on vendor instructions.

Transcription Queue Sweep: Installed by the Rms Add-on, the Transcription Queue Sweep has several configurable parameters to set. Depending on a customers requirements the following can be adjusted;

- scheduled to run during a specified time period
- maximum batch size of documents to process
- maximum threads to dispatch requests
- time between retry attempts for failed requests
- maximum number of times to retry a failed request

Technical Document: Transcription with FileNet Content Manager: Configuring Transcription and Building Transcription Handlers

http://www.ibm.com/support/docview.wss?uid=swg27043790

Course Roadmap

- Transcription Framework Overview
- Configuration
- ➔ Transcription Queue Sweep
- Demo: Configuration and Transcript Generation
- Best Practices
- Course Summary

Transcription Queue Sweep

- After the Rms Add-on has been installed, transcription document class and its subscription have been configured on an object store; transcription requests will be generated for newly checked-in video documents based on the filter expression that was configured in the subscription. The generated requests are stored in the Transcription Request Queue and processed by the Transcription Queue Sweep. Within the Queue Sweep a Transcription Request Sweep Handler:
  - Iterates through batches of requests
  - Submits the referenced document for each request to the Transcription Handler for conversion to TTML
  - Adds the TTML to the document as a Transcription Annotation

- The Transcription Requests generated by this process can be monitored (in ACCE) by opening the Transcription Request Sweep object under Sweep Management and selecting the Queue Entries tab.

Transcription Queue Sweep (cont.)

- An object store may already contain video documents, the content of those documents can be converted to transcripts also by a custom Transcription Queue Sweep Job. This is done with ACCE by creating a custom sweep job that:
  - Is a subclass of Transcription Job
  - Specifies the Transcription Document class as its target
  - Specifies a filter expression that limits conversion to video documents and/or defines a date range of the documents to be transcribed
  - Specifies when the sweep job will run

- The processing of the requests generated by the job is also be monitored by opening the Transcription Request Sweep object and selecting the Queue Entries tab.

Course Roadmap

- Transcription Framework Overview
- Configuration
- Transcription Queue Sweep
- ➔ Demo: Configuration and Transcript Generation
- Best Practices
- Course Summary

Demo: Configuration and Transcript Generation

The following demo will include the following transcription framework features:

- Location of the Rich Media Transcription Extensions Add-on
- Creating a transcription document class and its subscription
- New video content: checking in a video document, queuing of its transcription request and the transcript annotation created
- Existing video content: creation/execution of a Transcription Queue Sweep Job
- Associating the transcription document class with a Annotation Preprocessor
- Indexing of the transcript annotation and searching for the indexed text

Note: For development and testing of the CPE transcription framework feature a "test harness" was developed to simulate transcription processing. That test harness will be used in the following demonstration.

Course Roadmap

- Transcription Framework Overview
- Configuration
- Transcription Queue Sweep
- Demo: Configuration and Transcript Generation
- ➔ Best Practices
- Course Summary

Best Practices

- Recommended to create a separate transcription document class for rich media. This will not only allow for a separate storage location for these large documents but it will likely result in a significant performance improvement for indexing by limiting the execution of the transcription annotation preprocessor to only rich media documents.

- Because videos are large documents, it is more efficient to store them as files rather than in the database. So it is recommended that the default storage area for the transcription Document and Annotation classes is set to a file-based storage area.

Course Roadmap

- Transcription Framework Overview
- Configuration
- Transcription Queue Sweep
- Demo: Configuration and Transcript Generation
- Best Practices
- ➔ Course Summary

Course Summary

You have completed this course and can now:

- Describe the components of the Transcription Framework and how they interact to transcribe video documents and index the resulting transcript to provide full-text search capabilities of the video document.
- Have a basic understanding of the installation/configuration of the framework components.
- Understand how Transcription Queue Sweeps and Queue Sweep Jobs are used in the video transcription process.
- Know how to locate the technical document that provides in-depth configuration details.

Product Help/Documentation/Resources

Technical Document:

Transcription with FileNet Content Manager: Configuring Transcription and Building Transcription Handlers

http://www.ibm.com/support/docview.wss?uid=swg27043790

Contacts

- Product Manager: Stephen Hussey
- Development Manager: Grace Smith
- Subject Matter Experts: Shawn Waters, Brian Owings, Himanshu Shah, Antonio Montanana, Darren Kennedy & Eric Edeen