PUBLISHING PREPARATION SOLUTION

Overview Document

Copyright

Contact Information

Address:
2465 Roscomare Road
Los Angeles, CA 90077

Phone:
+1 310 740 3523
+1 310 776 9303

Fax:
+1 310 743 8067

Website:
http://www.dexik.com

Email:
support@dexik.com
sales@dexik.com

Table of Contents

Preface

Publishing Preparation Solution Overview Document describes the purpose and benefits of Dexik Publishing Preparation Solution and provides information on how to install and configure this solution.

Audience

The *Publishing Preparation Solution Overview Document* is intended for professional services, system integrators, business analysts and system administrators. The readers must have a good understanding of Windows system and general computer technology.

Chapter 1: Publishing Preparation Solution Introduction

This chapter introduces Publishing Preparation Solution and lists the system requirements.

The following topics are described in this chapter:

Product Overview

Publishing Preparation Solution is a powerful multi-tier tool designed to prepare large quantities of image files for publishing by wiping out the confidential information from the images. The confidential information can be identified by word masks or regular expressions, suitable for both free from documents and forms. The confidential information can also be identified by fixed position (templates); this approach works well only on the forms where the information maintains specific position. The template document must be the same type of form as the documents to be processed. It is possible to define multiple templates as well as multiple masks to be setup simultaneously.

The product allows using the specially designed QA station to verify the identification results before wiping the information out. The user performing QA can add, modify or delete the highlight that marks information to be hidden from the final, redacted document.

User requirements define the way how information from the image. Customer has a choice to choose between placing an annotation (for TIFF/DMS) over the sensitive data or burn a rectangle for permanent image modification. In the case, of permanent modification – it is recommended to keep the document original.

Customer can configure different ways how the documents are “triggered” to be processed for sensitive information identification and redaction. The solution can work of certain network share or database.

Mask identification requires installation and licensing of the ABBYY® FineReader® Engine 8.1 runtime. © 2007 ABBYY Software. ABBYY FineReader – the keenest eye in OCR.

The ABBYY® FineReader® Engine 8.1 runtime license must be acquired from local ABBYY representative through Dexik, Inc.

The system allows performing Publishing Preparation on multiple servers/workstations to achieve the desired overall processing speed. The actual performance depends on hardware and network capabilities.

The following image format is supported for Publishing Preparation:

VisiFLOW ™ DMS
TIFF

By request, the support could be extended for the following image formats:
PNG
BMP
PCX
DCX
JPG
JPEG2000

Why is Dexik Publishing Preparation Solution the best choice when it comes to automatic document redaction?

We believe Dexik Solution is the best for the following main reasons:

- Dexik Solution can be easily integrated with any Workflow and Document Management System.
- Dexik Solution setup and configuration is easy and mask setup require only basic knowledge of regular expressions.
- The QA Station component, which is part of the solution, provides an intuitive interface to effectively review the documents before redaction.
- Dexik Solution is able to utilize multiple workstations “on-demand” for QA purposes automatically. The user should just start the application and agree to QA.
- Dexik Solution allow customer to use multiple OCR Servers simultaneously to improve recognition performance when necessary.
- The unique processing algorithm allows indentifying “clue” data first and then looks for the data to redact, greatly improving document recognition and redaction accuracy.

System Requirements

This section lists Publishing Preparation Solution system requirements.

Publishing Preparation Solution Server System Requirements

Component	Description
Operating System	Windows 2000/2003/2008 Server (32-bit or 64-bit), Windows 2000/XP/Vista/7
RAM	2 GB or more
Available disk space	20 MB
Software	Microsoft .NET 3.5 SP1, ABBYY FineReader Engine 8.1 runtime.

Chapter 2: Publishing Preparation Solution Server Modules

This chapter reviews the processing logic and modules configuration on the server side.

The Publishing Preparation Solution package consists of four Server side modules:

- **AutoHide TG (Task Generator)** module is responsible for the input, output, and redaction masks and templates setup. This module is a single-point process configuration.
- **QA Manager** module is responsible tracking the Quality Assurance tasks.
- **OCR Processor** module is responsible performing Optical Character Recognition, based on the configurations provided by **AutoHide TG** module.
- **AutoHide Processor** is responsible for physically redacting the image, either by applying an annotation or burning a rectangle over the sensitive information identified by template or **OCR Processor** module.

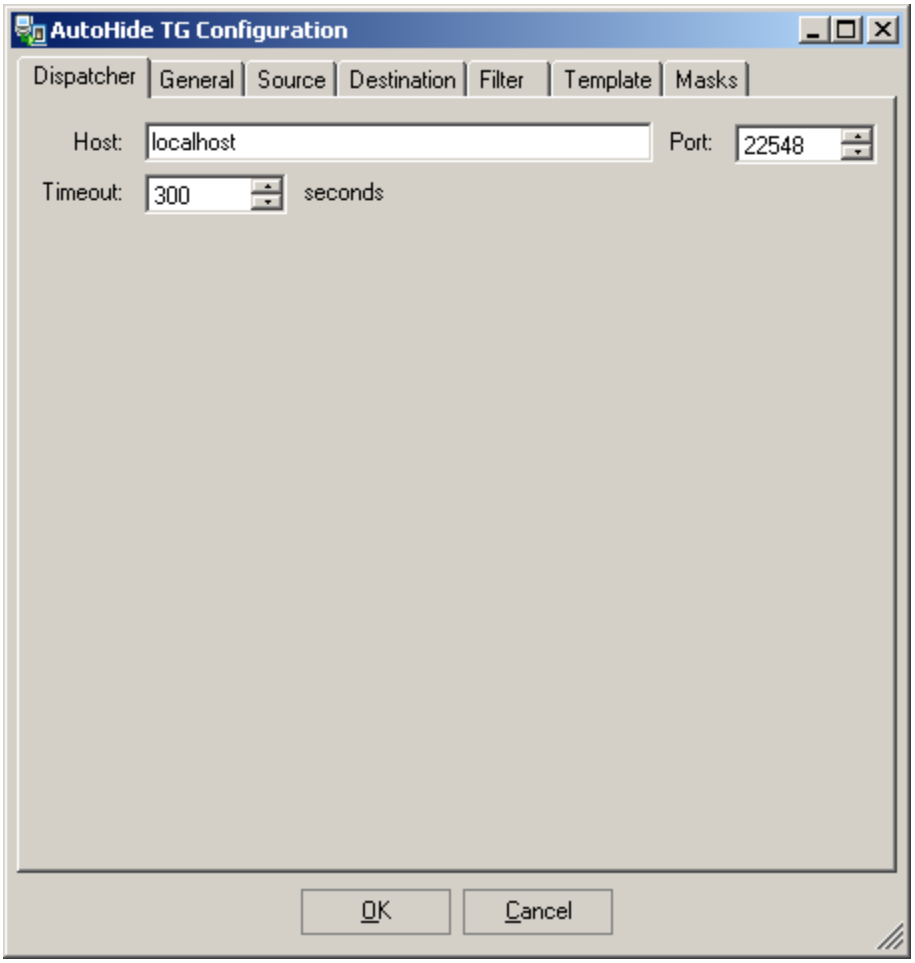
AutoHide Processor and **OCR Processor** modules do not contain configuration and are only used by the **AutoHide TG** module.

Here is how the solution works:

- Based on the setup, the **AutoHide TG** module determines the files that should be processed. The file list can be driven either by the physical files in the folder or by table in the database.
- Files are picked-up one by one and sent to **OCR Processor**, in case if the mask-driven configuration exists.
- If Quality Assurance option is selected, the recognized image should be sent to QA Station for user review. The **AutoHide TG** module submits the masks, identified data and the image to **QA Manager** module.
- **QA Manager** module submits information to one of the **QA Station** modules, installed on individual user workstations. The package is preserved for several reasons:
    1. User may not finish all the work and shutdown the PC;
    2. User may not be able to finish document review in time for whatever reason;

  The events above shouldn't stop the process and that is why if the timeout occurs on one of the packages, the **QA Manager** module will re-send the package to next available **QA Station**.
- User performs the QA on the recognition process using the **QA Station** UI and submits document further.
- The document arrives to the next step - the **AutoHide Processor**. This process creates annotation or burns out the data identified by OCR/templates or manually identified by the user during QA.

AutoHide Task Generator Module



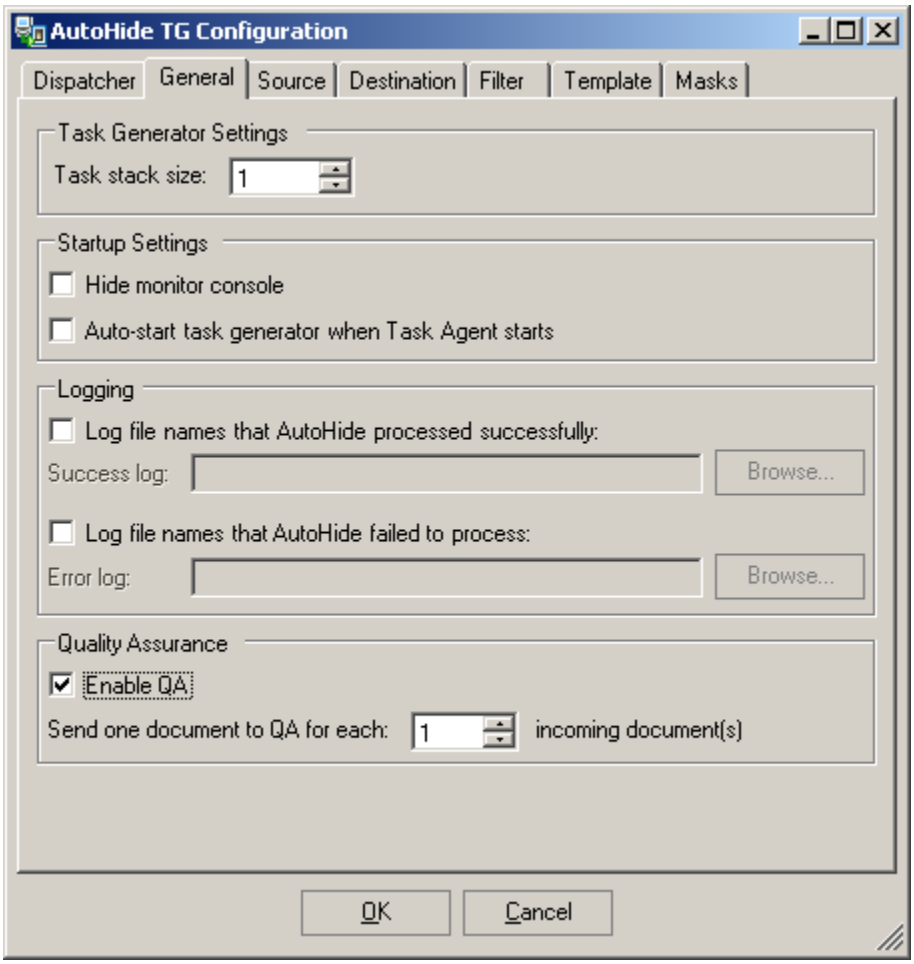
Dispatcher Configuration

This tab allows user to configure Dispatcher connection parameters: Dispatcher IP/host, connection port and task timeout. Task Timeout is the maximum time period reserved for task execution. If task is not executed within this period, the timeout error is reported for administrator's review. After that the system will try to reprocess the next file.

In a common case when default Dispatcher configuration hasn't been changed and AutoHideTG module resides on the Agent that's installed on the same machine as Dispatcher – the configuration shouldn't be changed.

General Configuration

This dialog allows to configure general processing parameters.



The **Task Stack Size** parameter represents the number of Publishing Preparation task requests that AutoHideTG server submits to Dispatcher for processing. This parameter should be used to tune the performance for multi-Agent processing systems.

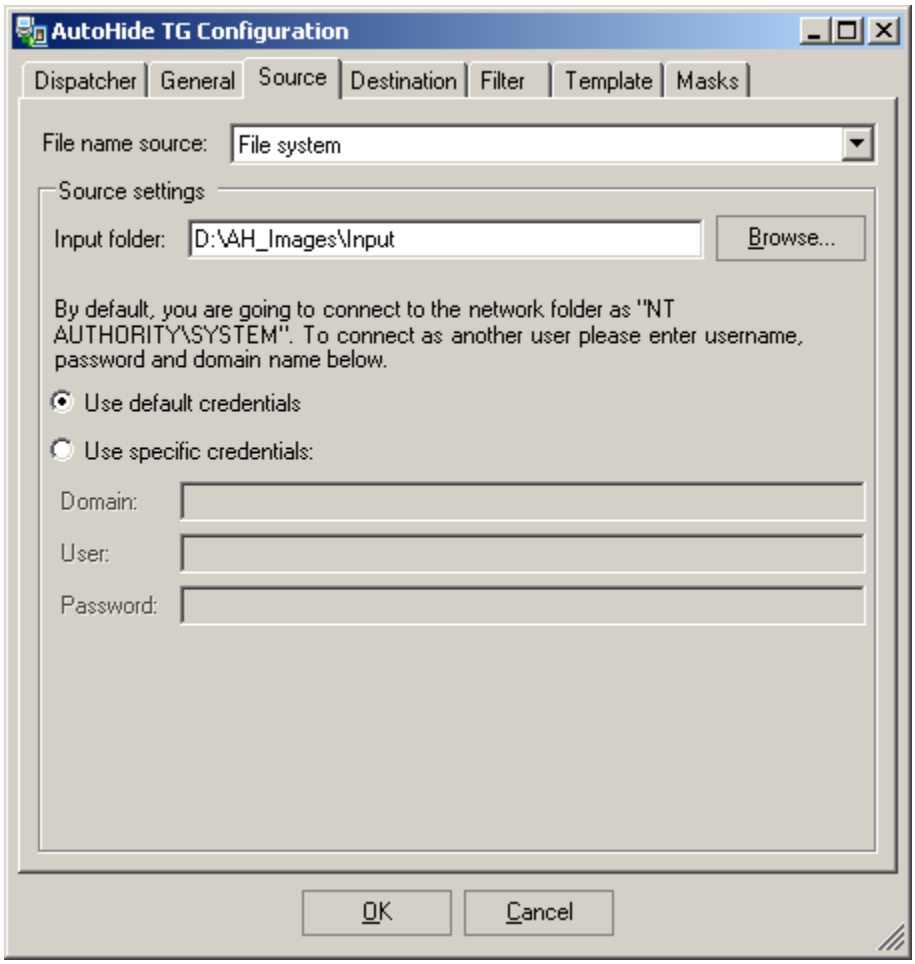
If **Auto-Start** option is selected, the AutoHideTG server will automatically start processing when Agent is started. This option should be selected if you do not wish to display Monitor Console and want to start Publishing Preparation processing.

Hide Monitor Console option allows hiding the AutoHideTG server Monitor console. By default, the Monitor Console is always shown.

Logging allows you to configure log file locations for both successfully processed files and the files that caused a processing error.

Source Configuration

This dialog is to configure source file location(s).



The processing source files could be picked-up for processing from either **File System** or **Database**.

In case if **File System** is a source for input files, the **Input Folder** and user access credential should be specified. In case if **Database** is used to determine the source file location, the database connection parameters, table and field names should be defined. The field should be a varchar-type field pointing to a specific network or local location.

Destination Configuration

This dialog allows configuring destination file location(s). It is similar to the **File System** option dialog for source configuration. There are only two options to choose from:

- Source files will be overwritten by the new files.
- New folder for the source files is selected.

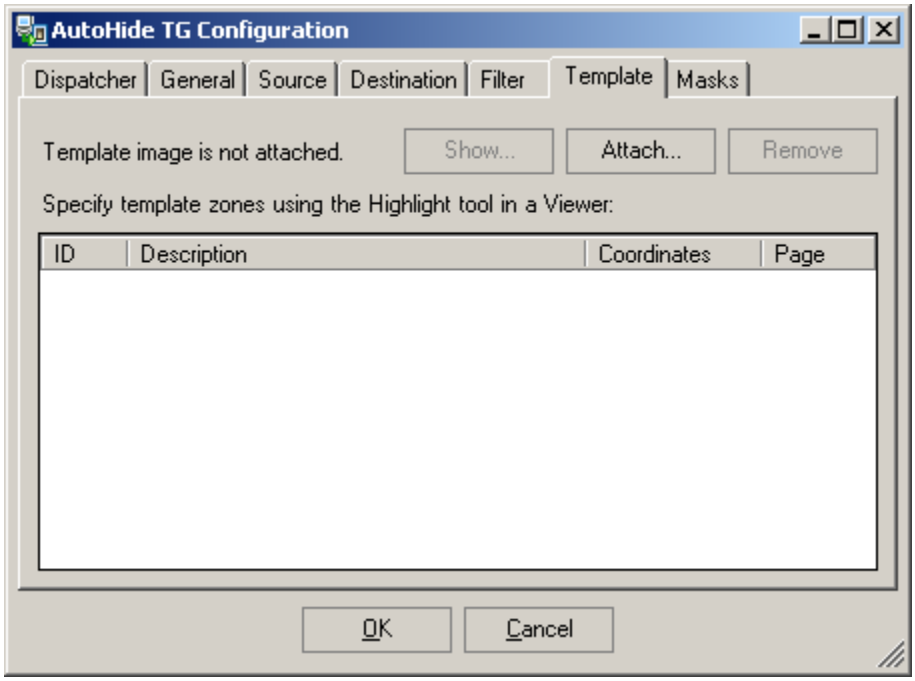
In case if new folder is used as destination location, the user access credentials should be specified.

Filter Configuration

This dialog allows configuring additional filter options for source files. Files could be filtered by filename mask or modification date(s). If filter option is defined – only the files that match the filter option will be picked-up for processing. In the opposite case, the application will try to process all files in the folder.

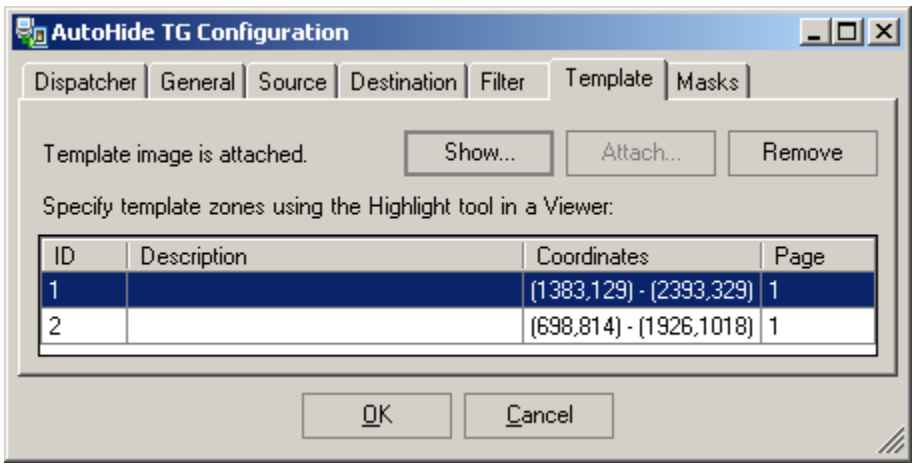
Template Configuration

This dialog tab allows attaching and defining image templates for form-like processing.



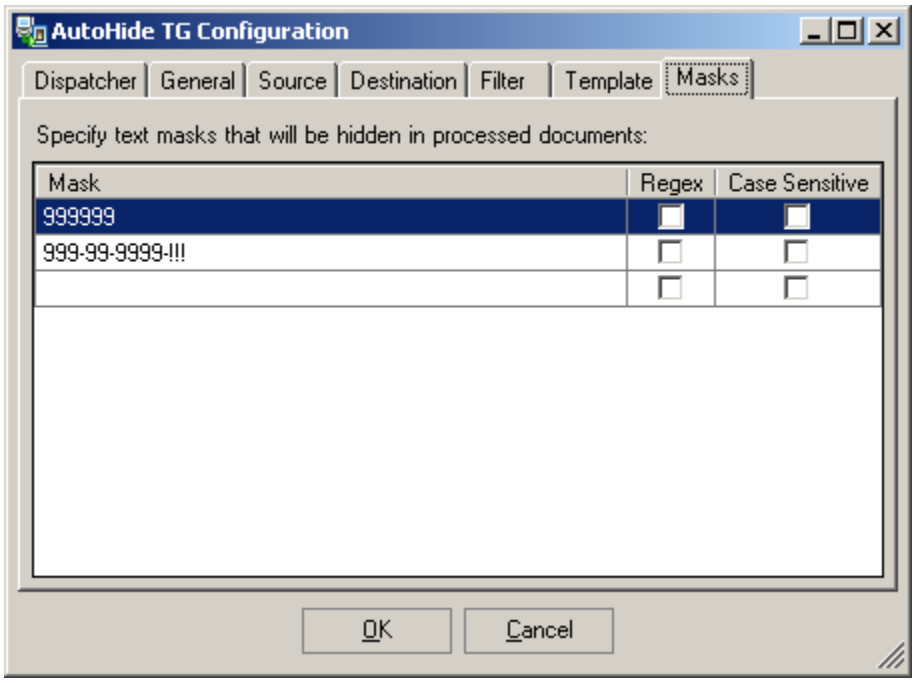
If no templates are defined, the **Attach…** button should be used to pick-up a TIFF file as a template.

After the template file is attached – please click **Show…** button, the DxViewer window will appear. Use the **Highlight** annotation to define zones and defined zones will appear in the Template window:



Masks Configuration

This dialog allows defining the masks “matches” for the free form document processing.



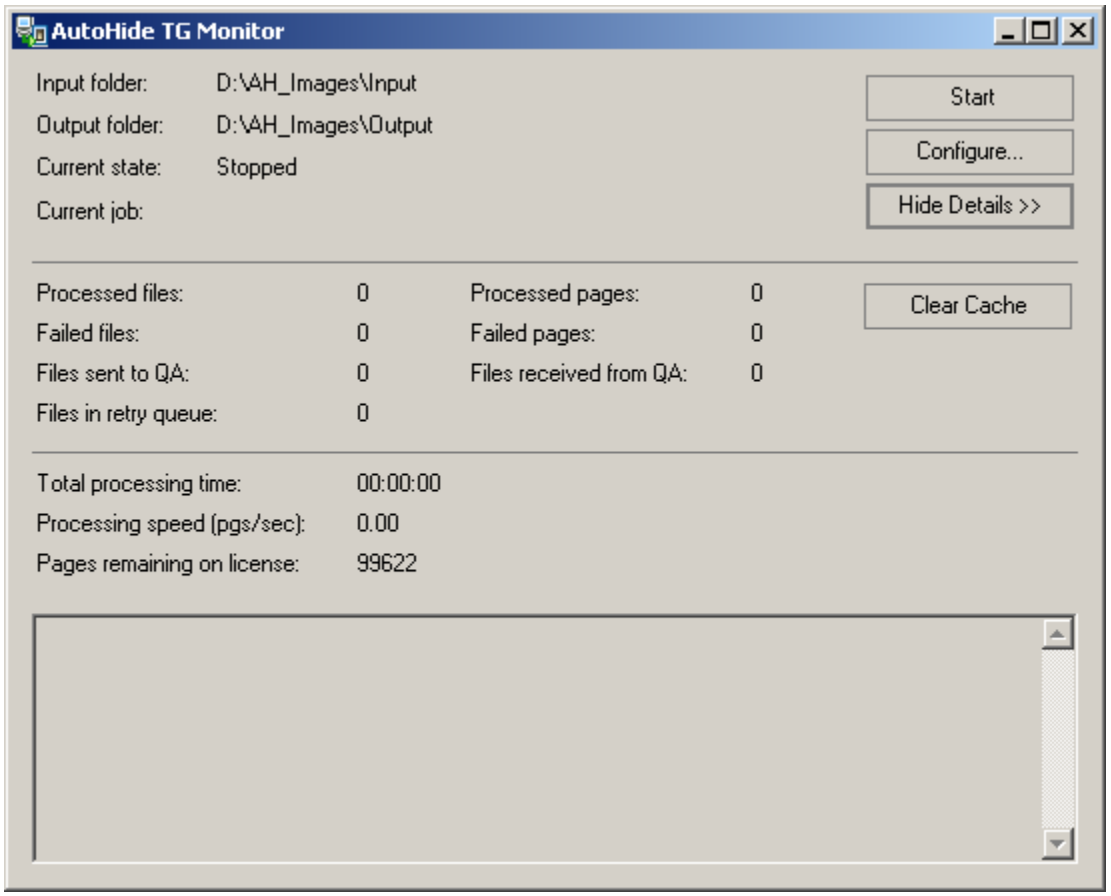
There could be multiple masks defined for processing and words matching each mask will be wiped out. The mask values are defined as follows:

- ! – any symbol;
- 9 – digits only;
- N – numeric;
- A – Alphanumeric.

Masks could be defined as regular masks or regular expressions and could be marked for case sensitive evaluation.

Publishing Preparation TG Monitor Window

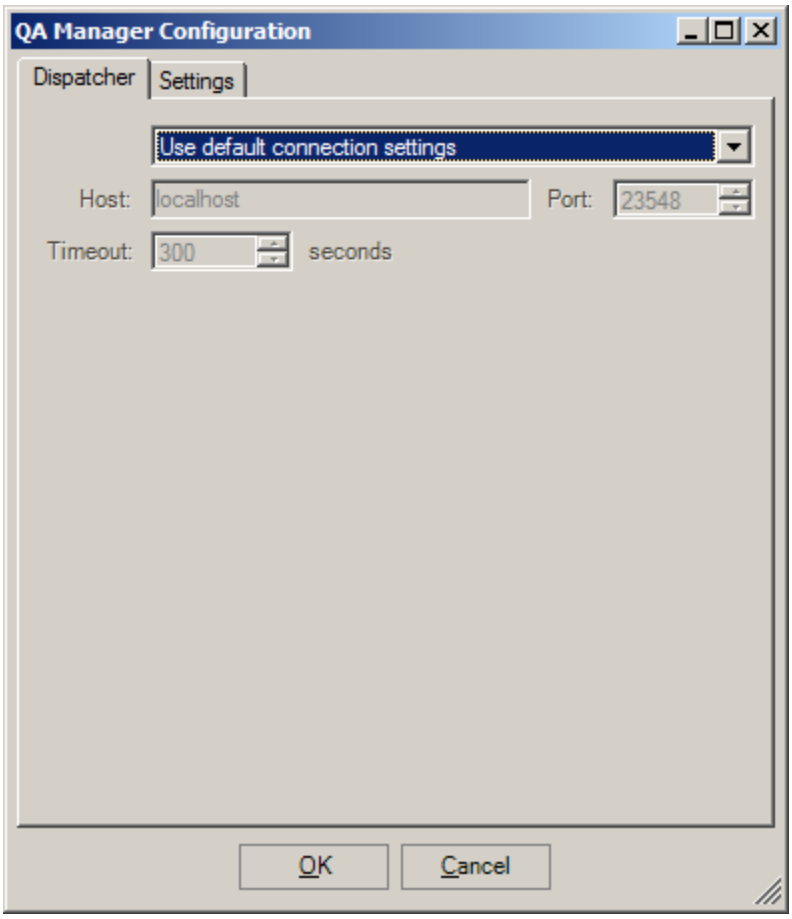
When Publishing Preparation TG Server is configured to show Monitor Console, the following window will appear on Agent startup:



This Console provides the ability to start/stop Publishing Preparation process, monitor Publishing Preparation progress and review statistics, like Processed Files Count, Processed Pages Count, Failed Files Count, Failed Pages Count, Files in Retry Queue, Pages Remaining on license, Total Processing time (from last time the process was started) and Average processing speed. Console allows accessing AutoHideTG server Configuration by clicking the **Configure** button.

The Publishing Preparation TG Monitor server caches the last file processed in order to avoid re-processing large amount of data in case of failure, therefore files should not be added to the folder when Publishing Preparation process was started and is not complete yet, this may cause some files to be left behind. In case if you need to re-process the all files in the selected folder please use **Clear Cache** button to allow reprocessing. Please be advised that reprocessed number of pages will be subtracted from the total number of pages licensed with Dexik.

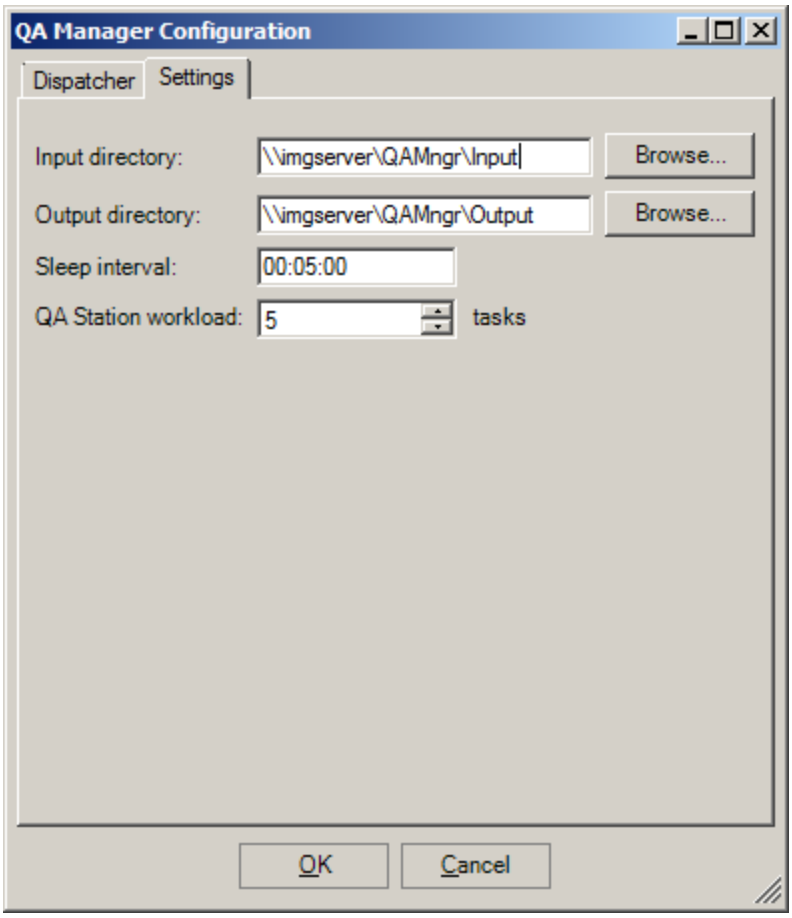
QA Manager Module



Dispatcher Configuration

This tab allows user to configure Dispatcher connection parameters: Dispatcher IP/host, connection port and task timeout. Task Timeout is the maximum time period reserved for task execution. If task is not executed within this period, the timeout error is reported for administrator's review. After that the system will try to reprocess the next file.

In a common case when default Dispatcher configuration hasn't been changed and AutoHideTG module resides on the Agent that's installed on the same machine as Dispatcher – the configuration shouldn't be changed.



QA Manager Settings

This tab allows user to configure QA Manager processing options.

The **Input Directory** is the file-system location, where **QA Manager** will store the files that haven't been sent to the **QA Station**.

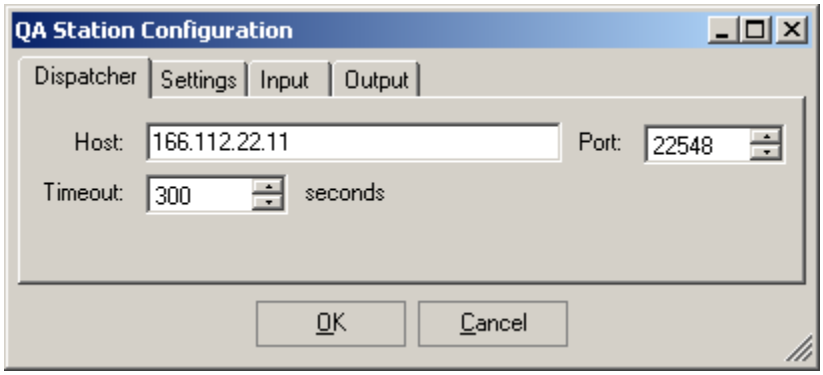
The **Output Directory** is the file-system location, where **QA Manager** will store the files that have been sent to the **QA Station**.

The **Sleep Interval** is a time period between job distributions from **QA Manager** to **QA Stations**.

The **QA Station Workload** is the parameter that controls the maximum number of tasks in the queue for each QA Station.

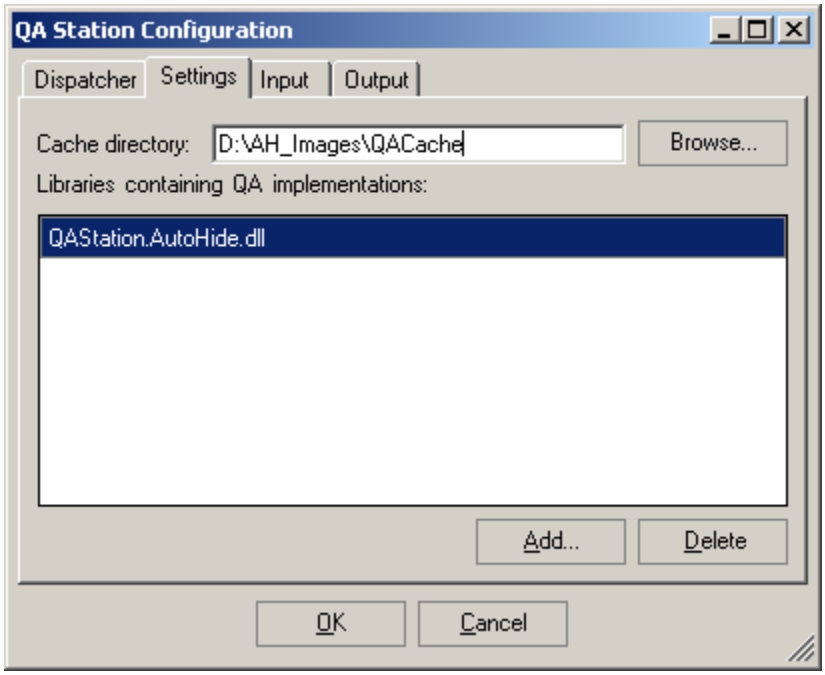
Chapter 3: Publishing Preparation Solution Quality Assurance (QA) Module

The Quality Assurance station purpose is to allow user verify the identifications made by the automatic process. It is designed to assist user with information review, speed-up the verification process and improve the overall system output quality and performance.

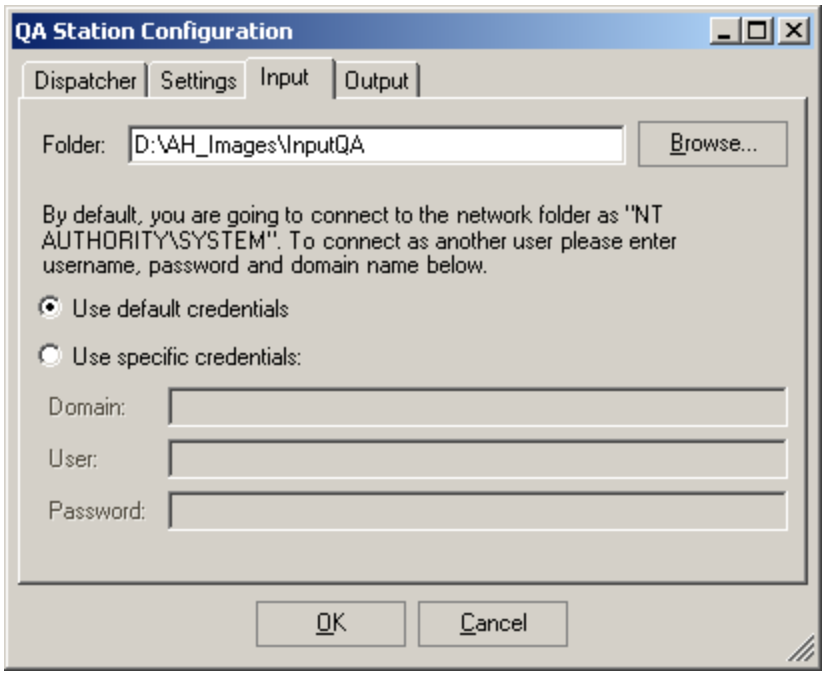


The **Dispatcher** tab allows configuring the connection to the Task Dispatcher application.

The **Settings** tab allows specifying the Cache directory for the QA module. It folder will be used to store temporary files.

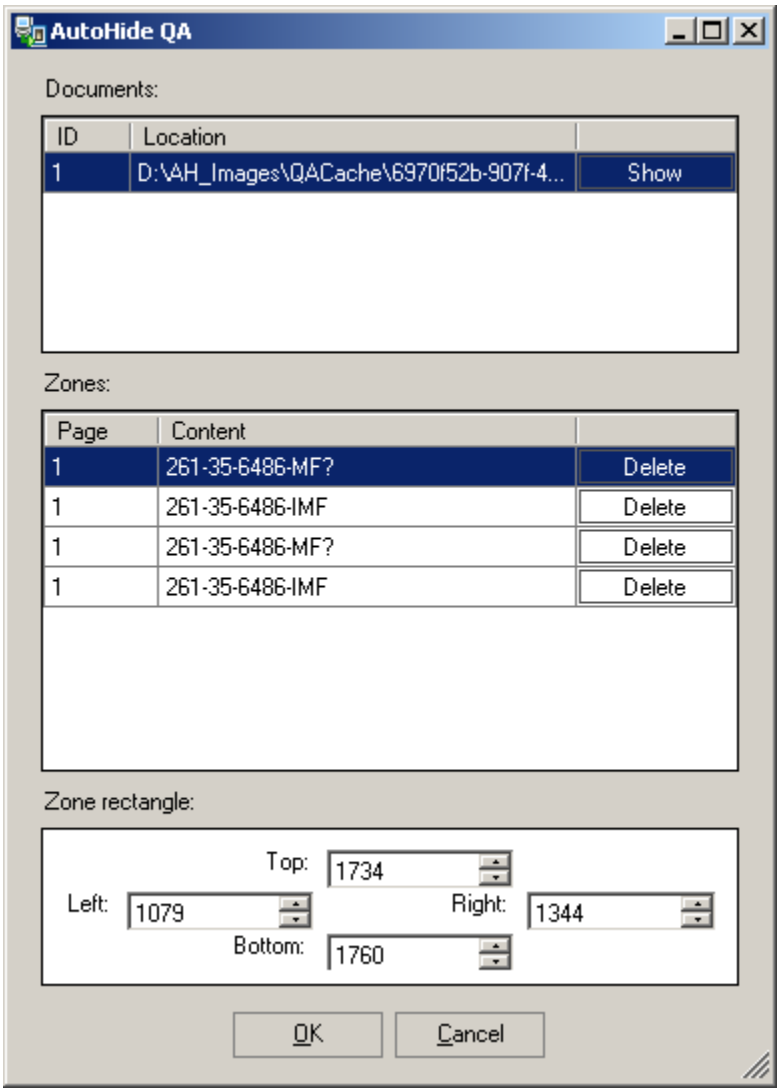


The **Input** and **Output** tabs allow specifying the input and output folders that the QA station will use during the processing.



When the QA station is fully configured it will be ready to receive documents for verification. If there are more than one QA stations currently connected to the system, the load balancing is performed by Dexik Task Agency, meaning the QA station with fewer documents for review will receive the next document.

When the document arrives for QA - the QA station window will appears.



User should click **Show** button and the **Dexik Viewer** window will open. The identified zones are highlighted on the image.

The user can use the QA station window to edit and delete existing zones. To add a new zone the user should use the Highlight tool in the viewer window. As soon as highlight is added on the image – the zone with empty content is added to the zone list.

Based on the corrected zones the server will later redact the information identified.

Useful Tips

Using mapped network drives with Dexik Publishing Preparation Solution

By default, Dexik Watch Dog component is run under the Local System account. It may cause an issue when working with the mapped network drives. The service under the Local System account usually is not able to work with the network drive. The Dexik Publishing Preparation Solution fully supports UNC path usage with the Windows user credential specification. In some exceptional cases, when UNC usage it not an option and the only possibility to access a network share is to map a network drive there are 2 possible solutions:

1. Configure DTS Watch Dog Service to run under specific user account in the Windows Services panel. There are following drawbacks for that solution:

Dexik Watch Dog, Dexik Task Agent and Dexik Task Dispatcher icons won't be visible in system tray and accessible through tray

Dexik Task Agent has to be configured via Task Agency Monitor only

AutoHide TG Monitor window won't be visible, limiting Publishing Preparation process monitoring

2. Run the DTS Watch Dog manually from the command line. To perform that please stop the DTS Watch Dog Service first and switch it to Manual starting mode. After stopping the service, in the command line type *"C:\Program Files\Common Files\Dexik\DTS\DSWatchDog.exe" /APP*. The only drawback to that solution is that in case of operating system restart, the Watch Dog and other applications will not be restarted automatically.