DIGIform
PO Box 390
Wilgeheuwel
1736
www.DIGIform.co.za
TEL: +27 11 675 7083
FAX: +27 (0)88 011 675 7083

COVERSHEET BASED DOCUMENT CAPTURE SOLUTION USING CARDIFF TELEFORM

WRITTEN BY: RICHARD DAVIS

DATE: 9 OCTOBER 2007

TABLE OF CONTENTS

1 Overview

1.1 Document Solution Required

The client received a large number of variable layout and low quality identity documents and other, unstructured, supporting paper documentation.

The requirement was to assign a document class and index all the documents in the quickest, most accurate and most automated manner possible.

After processing the documents would be exported to the client's content management system.

Identification documents are required to be stored in a format that is optimised for picture recognition and size. Supporting documentation is required to be stored in a format that is optimised for text visibility and size.

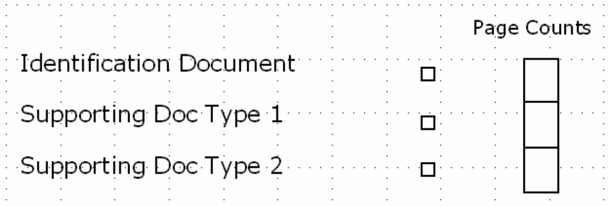
This document serves to outline the system implemented.

2 DETAILS

2.1 COVER SHEET DETAILS

A coversheet was designed within TELEform as a Traditional TELEForm Form. The main fields on the cover sheet were as follows:

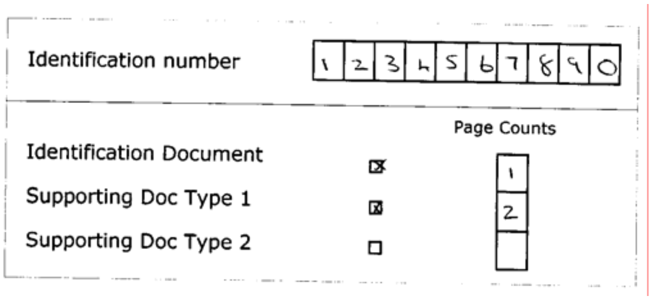
- A constrained print field to capture the main index criteria (the applicant's identification number).
- Check boxes to indicate the presence of the various types of supporting documents.
- A single constrained print field to capture the number of pages of the relevant supporting document types
- A static barcode whose value allowed identification of the coversheets types and version.



		Page Counts
Identification Document	□	
Supporting Doc Type 1	□	
Supporting Doc Type 2	□	

2.2 DOCUMENT PREPARATION

As part of the document preparation the cover sheets are completed, by hand, with the relevant information and then the referenced supporting document attached in the order and numbers indicated.



Identification number 1 2 3 4 5 6 7 8 9 0

		Page Counts
Identification Document	☒	1
Supporting Doc Type 1	☒	2
Supporting Doc Type 2	□	

These documents are then placed in boxes pending scanning.

2.3 DOCUMENT SCANNING

Documents are bulk scanned in a high-speed scanner in batches of up to 300 pages.

Scanning is performed via DIGIform's ScanBurst module. All documents are scanned in greyscale to avoid identification picture quality loss.

As a configurable option ScanBurst can use converted monochrome documents to speed up processing while keeping a copy of the original high quality greyscale images for final merge and export.

ScanBurst bursts the documents based on the Bar-coded value on the Coversheet.

The benefits of splitting the documents into individual batches within TELEform are as follows:

- Prevents very large batches and the associated space and performance issues when using the scanner capacity to its fullest.
- Avoids large batches needing to go for review if some of the logical document packs within the batch have incorrect or low confidence page count indicators.
- Allows for very fast pass through of batches to the backend content management systems.
- Allows for granular tracking and monitoring of batches.

Scanburst bursts and processes the scanned batches asynchronously. i.e. the scan user can continue to scan regardless of the speed of the backend conversion process. This is especially important when a high-speed scanner is used and documents are scanned at a high resolution and / or in greyscale or colour.

In addition a ScanBurst server side component can be installed to take over some of the processor intensive splitting functions such as Barcode detection and image manipulation.

Once the document has been processed it is exported to TELEform via TELEform's Auto-Batch-Create folder (using batch header files) pickup functionality or via ScanBurst's "TELEform Direct" connect agent. The "TELEform Direct" connect agent avoids potential delays associated with document picking up, and validation, associated with traditional Auto-Batch-Create process.

2.4 TELEFORM PROCESSING

TELEform receives the documents from ScanBurst as a greyscale batch. Each batch matches a logical document set and consists of between 2 and 6 pages. (It is possible for the TELEform system to process multiple document sets within one batch, which could occur if a splitter barcode were not recognised)

TELEform evaluates the coversheet and reads the handwritten identification number, check boxes and page number indicators. It then performs a number of processes:

- Validates the identification number using a predefined algorithm
- Validates the number of pages indicated on the coversheet against the actual number of pages present in the current document set
- Checks for low confidence OCR character results
- Performs throughput enhancing checks such as clearing positive checkbox values that have no corresponding page count values

If any of the criteria above fail the batch is held for manual user inspection and correction. Otherwise the documents are split into their relevant document classes, based on the pages indicated on the coversheet, and exported to a backend holding database.

2.5 Document Compression and Content Management System Export

2.5.1 Compression Module

Once in the backend holding database the documents are compressed to predefined (and configurable) image formats.

In general the identification documents would be compressed to greyscale “Jpeg” compression and the supporting documents to a monochrome “Tiff Group 4” format.

Identification document image sizes are generally between 30 and 60 kilobytes while maintaining clear photo visibility.

2.5.2 Export Module

The final step in the process is the export of the various split documents, and their metadata, to the content management system.

The export module can export up to 10 documents a second and features enterprise capabilities such as automatic (and configurable) retries and report facilitating date tracking.

2.6 Throughput and Accuracy

The average time for documents not requiring user intervention (verification) from the time they were scanned till the time they appeared in the content management system ranges from 1 to 5 minutes (depending on volumes being experienced). The images in the content management system were in a final state i.e. document classed, indexed and compressed to the document types predefined compression settings.

On average between 20,000 and 25,000 pages are processed in an 8 hour shift. Of these more than 95% of the documents do not require user intervention once scanned.