Workflow Best Practices
3 examples

by
Maximilian Kleiss
Wolfgang Novak
Stephan Tratter
TREVENTUS - Overview

- **ScanRobot® 2.0 MDS**
  Automatic Book Scanner

- **ScanGate™**
  Image Processing

- **ScanFlow™**
  Workflow Management

- **Nainuwa™**
  Digital Library
Introduction I – The idea

Physical Material → Digital Library
Introduction II – Scanning

Physical Material → scanning → Digital Library
Introduction III – Digitizing

- Inhomogenous
- Exceptions
- Staff
- Preparation
- Scanning
- Quality control
- Physical Material
  →
  Digitizing
  →
  Digital Library
- Exceptions
- Output formats
- Storage
- Archiving
- Ocr
- Metadata
- Languages
- Exceptions
Workflow

Physical Material → digitizing → Digital Library

- inhomogenous exceptions
- scanning exceptions
- quality control exceptions
- staff preparation

Digital Library

- exceptions
- storage
- archiving

- output formats exceptions
- metadata exceptions
- OCR exceptions
- languages archiving
Workflow: examples

Complexity depends on:
- Desired end results
- Required quality
Workflow: examples

Complexity depends on:
- Desired end results
- Required quality

→ Simple workflows:
ScanFlow™

Best Practice:
Example A
„Law Files“
**Example 1: Material / Documents**

<table>
<thead>
<tr>
<th>BATCH</th>
<th>LAW FILE</th>
<th>CATEGORY</th>
<th>Single Sheets</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Folder" /></td>
<td><img src="image2.png" alt="Folder" /></td>
<td><img src="image3.png" alt="File" /></td>
<td><img src="image4.png" alt="Documents" /></td>
</tr>
<tr>
<td>000001</td>
<td>BAX120304</td>
<td>BAX120304-01</td>
<td>BAX120304-01_001.ext</td>
</tr>
</tbody>
</table>

- Total pages: ca. 1,5 Mio. pages (pilot project)
Example 1: Requested end result

- 1 physical page
  - 1 MASTER file – TIFF
  - 1 DERIVATE file – JPG
  - 1 file searchable – PDF

<table>
<thead>
<tr>
<th>RAW - SCAN</th>
<th>CROP-1</th>
<th>DESKEW</th>
<th>CROP-2</th>
<th>Extrapolate &amp; Background Homogenization</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASTER</td>
<td></td>
<td></td>
<td></td>
<td>DERIVATE</td>
</tr>
</tbody>
</table>
Example 1: Infrastructure
Example 1: Challenges

- Sheet Feeder Scanner output
  - Convert to target structure
  - Quality control: Completeness check
  - Renaming based on metadata from clients DB
- Synchronization with clients DMS (handshake)
  - Finished ScanJobs
  - Finished batches
- Etc.
Example 1: File & Folder - Renaming

1. **Scanning (sheet feeder)**

2. **Data and Workflow Server**

3. **SCANNER**

4. **WORKFLOW**

5. **IMPORT & CONVERSION**

6. **RENAMING**

7. **1 BATCH**

**XML Code**

```xml
<Batch-ID:15>
  ID_01=BAX030406
  ID_02=BAX030407
  ID_03=BAX030412
  ...
  ID_10=BAX030428
</Batch-ID:15>
```
Example 1: Synchronization w. extern SW

Sheet Feeder | ScanFlow | DMS Software (client)
---|---|---

000001 | 000001 | BAX030406

Rename:
+) Folders
+) Files

Batch is finished!

Batch-end.log
Example 1: Numbers and Figures

- Project time: 12 months
- Full throughput: ca. 120,000 pages/month
  - Including: scanning
  - Image treatment
  - Various Quality control steps
ScanFlow™

Best Practice:
Example B

“Register archive – QR codes”
Example 2: Material / Documents

Project size:
Register books > 6,000
Pages > 2 Mio
Example 2: Requirements

<table>
<thead>
<tr>
<th>One register consists of:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colored &amp; grey pages</td>
</tr>
<tr>
<td>Empty pages</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Maps</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Page count &amp; File naming</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Customer wants the whole work-flow steered by QR codes.
Example 2: Requested end result

Requested result:
- Table of Content: PDF searchable
- Chapter wise: 1 PDF per chapter
- Single pages: JPEG
Example 2: Infrastructure

General overview: Digitization Project „BARCODES / QR-Codes“

<table>
<thead>
<tr>
<th>Staff &amp; SW</th>
<th>Nr.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Solution:
- Developing a QR code system with small codes (<1 x 1cm) → no information lost
- Spooler function in QR code generator → avoids half empty pages → saves money

Using QR-Codes for steering
→ CQ-Codes are stuck on the pages where sequence changes:
  • start and stop scanner,
  • completeness check by page count information in QR code,
  • meta data (page content) added from QR code
  • place holder added,…
Example 2: qrCodeGenerator
Example 2: Challenges

Challenges:
1. Work-flow and scanner steered by QR codes.
2. If multiple QR Codes on page → detect correct QR code.
3. Different scanners → SW sorts the scanned pages automatically.
4. QR codes → as small as possible
5. QR codes → information about
   a. image treatment
   b. job handling

QR code

Scanner steering
Meta data
Page information
Processing information

Planetary scanner

map

ScanFlow™

L R

L R

L R

map

L R
ScanFlow™

Best Practice:
Example 3

„Dissertations / Diploma Thesis“
Example 3: Material & basic conditions

- 5000 Dissertations & diploma thesis 3-4 times per year
- ~150 pages/thesis
- Format: A4 with exceptions
- Thesis with simplex and duplex print
- Content: mainly text but also images, diagrams, tables
- ~ 2.5 Million pages/year; ~ 750k per subset
- Digitization of each subset within 6-8 weeks
Example 3: Infrastructure

General overview: Digitization Project „Dissertations / Diploma Thesis“

<table>
<thead>
<tr>
<th>Staff &amp; SW</th>
<th>Nr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- ScanRobot® 2.0 MDS
- IPP & QC
- IPP & QC (Final & RELEASE)
- Data and Workflow-Server
- OCR
- DMS

Consulting by TRENTUSConsulting GmbH
Project: Digitization Project „Diploma Thesis“ - Workflow v03
Creation: 10.07.2014
Last Update: 16.07.2014
Example 3: Requested end result (selection)

✓ Resolution: 400 dpi
✓ Cropping
✓ Scaling
✓ Naming convention (00001t.tif, 00002.tif)
✓ Removing blank pages
✓ File format: TIFF (CCITT group 4 & LZW compressed)
✓ OCR for abstract only
Example 3: Challenges & Exceptions I

- Deleting blank pages

![Workflow Diagram]

Page content:
- Not defined: 365
- Text: 0
- Image: 0
- Text and image: 0
- Table: 0
- Empty page: 54
Example 3: Challenges & Exceptions II

- File format: TIFFs depending on content

Workflow:

- Colour & Greyscale
  - TIFF
  - CCITT group 4
- Black & White
  - TIFF
  - LZW compressed
- Colour & Greyscale
  - Black & White
  - TIFF
  - CCITT group 4
  - LZW compressed
Example 3: Challenges & Exceptions III

☑ OCR for abstract pages only

Workflow

Abstract

TIFF

CCITT

group 4

TIFF

LZW
compressed

Meta data
workflow

CONCLUSIONS
Conclusio: Digitization Workflow characteristics

- Embraces **all areas** of digitization (Metadata, OCR, QC, final results, DB handshakes,…)
- **structures the flow** of work from the book to the final result
- **reduces communication** overhead ("He, does anyone know if the missing pages of this big brown book has already been rescanned?")
- enables you **define quality control** steps according to your needs
- enables you to **split the work** to subtasks and distribute it according to your resources
- enables you to **make changes** to all areas of your workflow at any time (exeptions, unexpected challenges)
- gives you an **overview and status** of all of your digitizing projects
Conclusio: Digitization Workflow

To put it in a nutshell:

"A good solution for your digitization projects makes your life easier, reduces the required resources and helps to increase the output in terms of quality and quantity."
Thank you!

www.treventus.com
tratter@treventus.com