

Dokumentenerfassung und -indizierung[1]

Dozent: Prof. Dr. Michael Eichberg
Kontakt: michael.eichberg@dhw.de, Raum 149B
Version: 1.0.1

Folien: [HTML] <https://delors.github.io/dm-erfassung-und-indizierung/folien.de.rst.html>
[PDF] <https://delors.github.io/dm-erfassung-und-indizierung/folien.de.rst.html.pdf>
Fehler melden: <https://github.com/Delors/delors.github.io/issues>

- [1] Dieser Foliensatz basiert auf Folien von: Klaus Götzer.
Dokumenten-Management von *Klaus Götzer, Patrick Maué, und Ulrich Emmert*,
dpunkt.verlag, 2023.
Alle Fehler sind meine eigenen.

1. Quellen von Dokumenten

Quellen von Dokumenten - Dimensionen

- Eigenerstellte und fremderstellte Dokumente
- Papierdokumente und elektronische Dokumente
- Einmalige Übernahme und laufende Übernahme

Eigenerstellte Dokumente

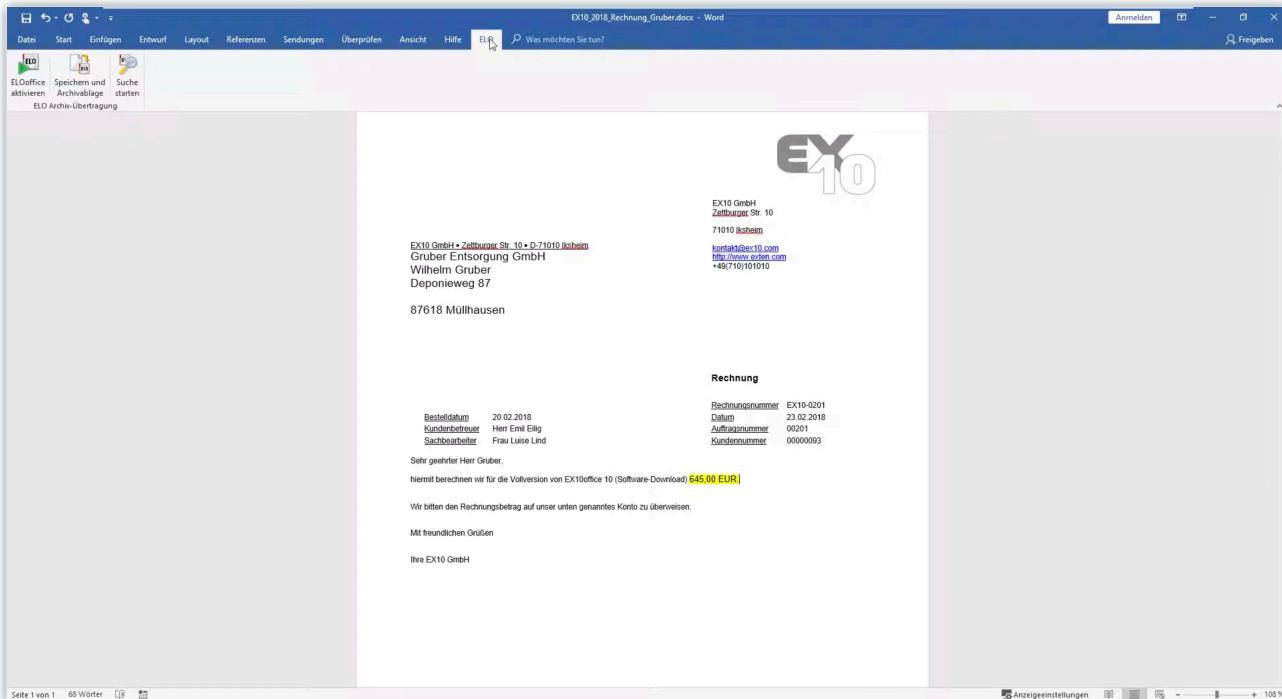
- Editoren für Texte, Graphiken, Mails, etc.
(Office, Outlook, AutoCAD,)
- Dokumentenerzeugende Systeme
(z. B. Rechnungen aus ERP-Systemen) (COLD)
- Übernahme von Bildern aus speziellen Verfahren wie Röntgen



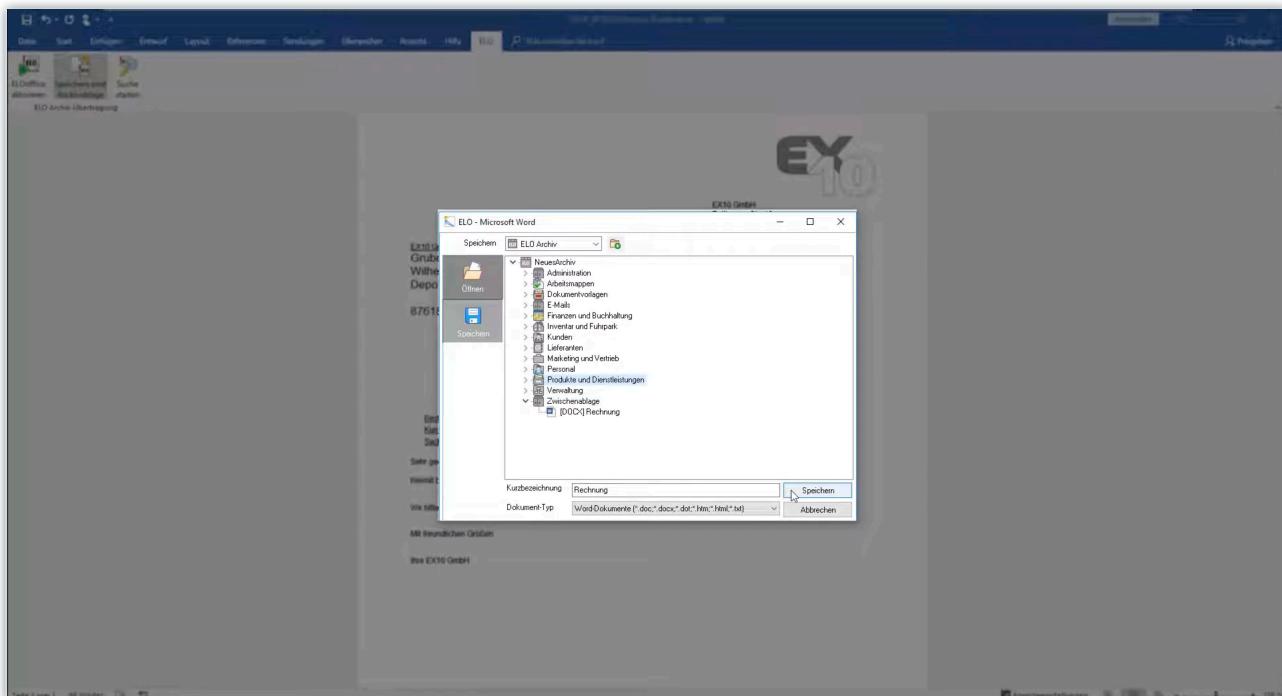
Bewertung

Anzustreben ist, dass beim Speichern automatisch Dokumente und Metadaten der Dokumente in das DMS übernommen werden.

Integration mit Office-Anwendungen



Speichern von Dokumenten aus Anwendungen (hier Word)



Fremderstellte Dokumente

Herkunft der Dokumente

- Posteingang (Papier)
- Übersendete Dateien
- E-Mail-Eingang

Typische Problemstellungen

- Unterschiedliche Formate
- Ermittlung und Erfassung der Metadaten

Probleme beim Eingang als Papier

- Aufbereitung des Eingangs
- Qualitätsunterschiede
- Umsetzung in ein CI-Format

NCI: *Non Coded Information* (z. B. Texte in Bildern)

CI: *Coded Information*

„Analoge“ (NCI) oder elektronische(CI) Dokumente

Papierdokument

- S/W oder farbig?
- Automatisch auszuwerten?
- Aufwand für manuelle Vorbereitung (Entheften, Glätten, ..)

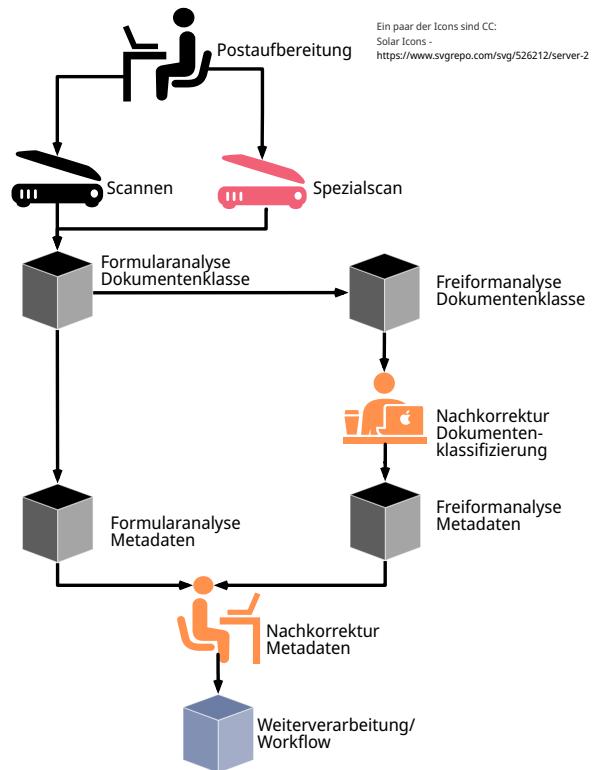
Elektronische Dokumente

- Welches Dateiformat liegt vor? Konvertieren?
- Automatisch auswertbar?

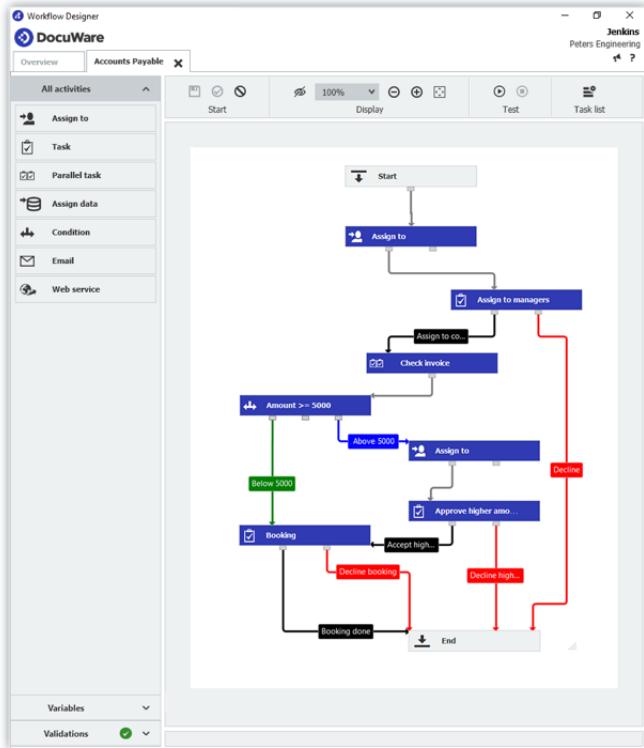
Strukturiertes Dokument oder Fließtext?

Beispiel für Eingangspostbearbeitung

- Workflow zur strukturierten Abarbeitung
- Ausnahmebehandlungen vorsehen
- Möglichst automatische Klassifikation und Indizierung



Unterstützung für Workflowdefinitionen in ECM Systemen - z. B. DocuWare



ECM: *Enterprise Content Management*

Erstmalige Übernahme von Dokumenten

Quellen

- Altsystem (Archiv, DMS)
- Filesystem
- Mikrofilm, Mikrofiche etc.
- Papierbeständen

Zu Klären

- Was ist wirklich sinnvoll zu übernehmen?
- Automatisierbare Übernahme möglich? (Zeitaufwand!)
- Outsourcing prüfen

Laufende Übernahme

- Eingehende Papierpost
- Eingehende E-Mails
- Ausgehende Dokumente
- Ausgehende E-Mails
- Fortschreibungen von Dokumentationen, Akten etc.



Bewertung

Zentrale Aspekte

- Etablierter „revisionssicherer“ Prozess
- Möglichst „Vollautomatik“

Automatisierung des Posteinganges (Papier)

■ **Sichere Übernahme des Dokuments in das DMS/Archiv**

- Protokollieren des Eingangs
- Zählen (Scanprozess) und paginieren
- Zeitsignatur / Bearbeitersignatur

■ **Klassifikation des Dokuments und Indizierung**

- Manuell durch Bearbeiter
- Automatisch (Formularerkennung, OCR - Volltext, Barcode)
- Gemischte Verfahren

■ **Zuordnung zu einem Geschäftsvorfall**

- Abgeleitet aus Metadaten
- Durch Bearbeiter

■ **Weitere Bearbeitung veranlassen**

- Weiterleitung (E-Mail)
- Workflow

2. Scanning von Dokumenten

Scannen der Eingangspost

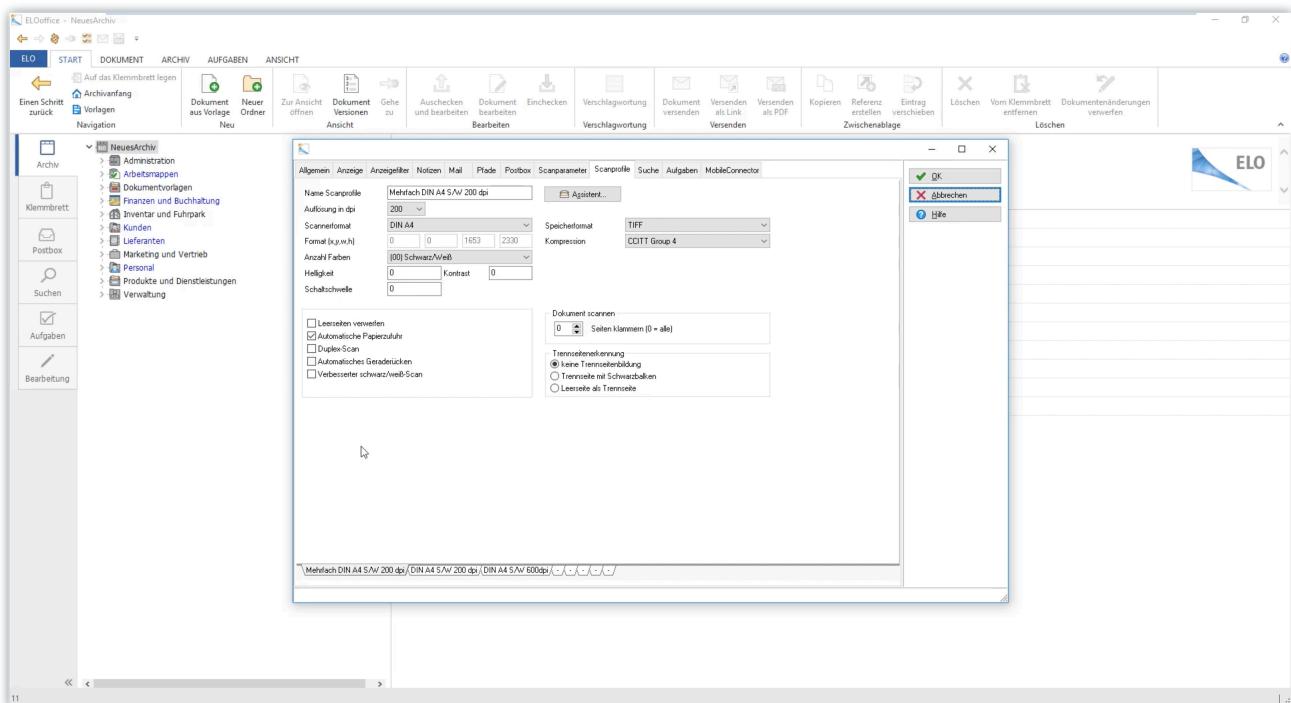
- **Scanner** sind die gängigsten Erfassungsgeräte für Dokumente auf Papier oder Film

Prozess

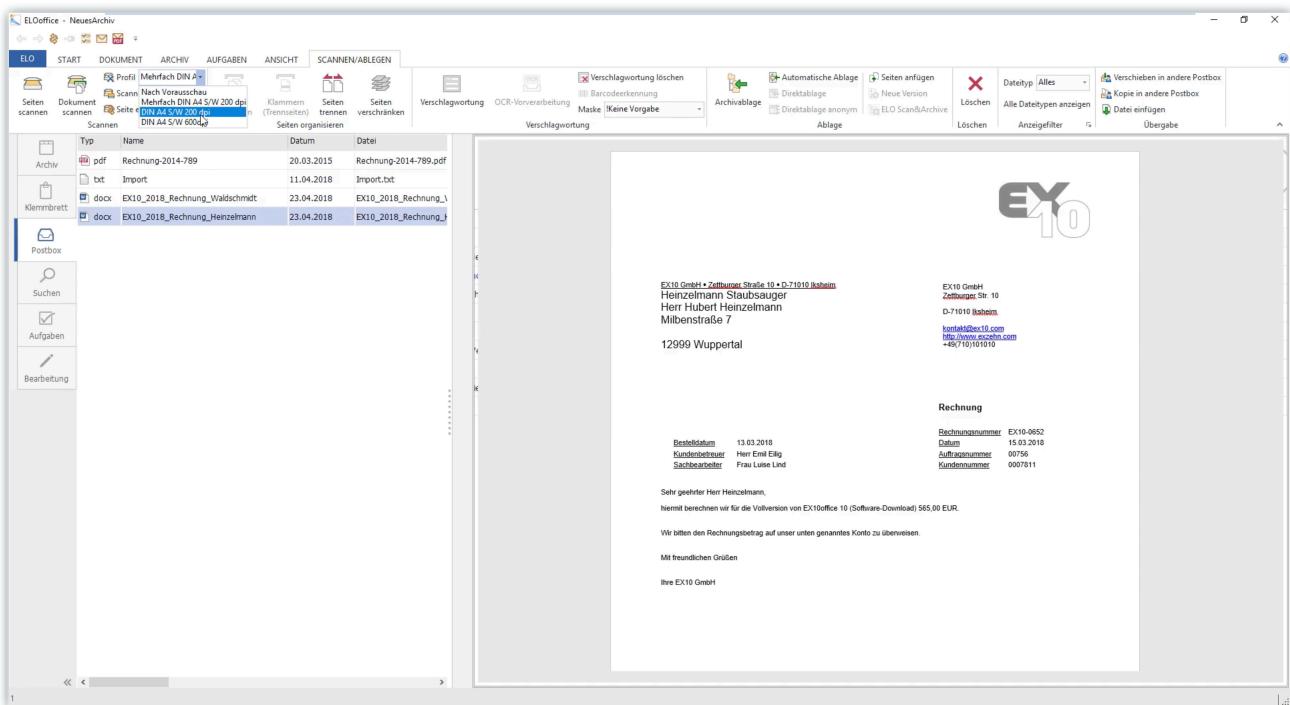
Papierdokument → Scannen → Elektronisches Dokument

- Scanning ist ein komplexer mehrstufiger Prozess zur Erfassung von Dokumenten
- Scanning ist meist mit weiteren Verarbeitungsschritten eng verknüpft.
- Zum Scannen und der Folgebearbeitung werden oft Speziallösungen eingesetzt.

Scanprofile (hier in Elo Office)



Vordefiniertes Scanprofile (hier in Elo Office)



Festgelegt wird:

- Auflösung
- Farbe oder S/W
- Trennseiten
- Barcodes
- Duplex
- Zielformat
- ...

Scanner

Scanner unterscheiden sich in:

- Zufuhr von Seiten
- Vorlagengröße (z. B. A4, A3)
- Geschwindigkeit (bis zu mehrere hundert Seiten pro Minute)
- Farbtiefe
- Umschlagerkennung
- Heftklammererkennung
- Preis
- ...



Scanmachine

Weiterverarbeitung gescannter Dokumente

- Umwandlung von Images (NCI) im CI-Dokumente (wie Texte)
- Klassifikation und Indizierung der Dokumente
 - manuell
 - automatisch
- Automatisches Auslesen von Formulardaten
- Automatisches Auslesen von Rechnungen oder ähnlichem
 - (Z. B. wenn die Dokumentenklasse bekannt ist.)

Umwandlung von NCI zu CI

Optical Character Recognition (OCR):

Primär auf Basis der Form der Zeichen der Maschinenschrift werden Pixelmuster in Zeichen umgesetzt.

Handprint Character Recognition (HCR):

Erkennen von handschriftlichen Texten.

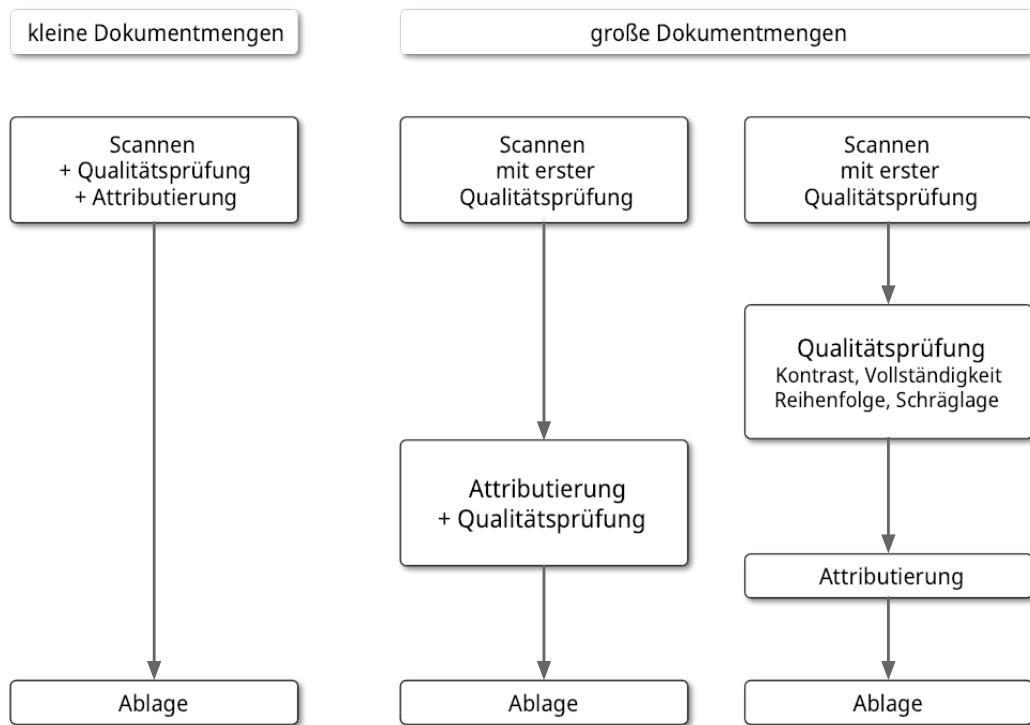
Intelligent Character Recognition (ICR):

Weiterentwicklung von OCR und HCR: Das Ergebnis wird verbessert durch modernste Algorithmen und KI-Verfahren.

Optical Mark Recognition (OMR):

Es werden Markierungen in vordefinierten Feldern/Bereichen ausgelesen. Z. B. Selektionsfelder aus Fragebögen oder es wird geprüft, ob „eine Unterschrift“ in dem vorgesehenen Feld erfolgt ist.

Arbeitsablauf beim Scannen



Sicherstellung der Qualität

Fehleranzahl hängt stark ab von...

- Vorlagenqualität (Knicke, Schmutz, ...)
- Schriftgröße
- Sonderzeichen
- Schriftart (mit/ohne Serifen...) und Qualität des Ausdrucks
- Qualität der Software
- Vorinformationen (welche Schriftarten werden verwendet...)

Problemfälle

- Ligaturen (z. B. „ff“ statt ffi oder „fi“ statt fi)
- Bestimmte Zeichenkombinationen z. B. rn: „r“ gefolgt von „n“ oder „m“
- Großes I (wie Ida) und kleines l (wie lieb) bei serifenlosen Zeichensätzen
- Fremdsprachige Zeichen (z. B. „\$“, „¥“ oder „£“)
- Optisch beschädigte Zeichen

Es muss **unterschieden werden** zwischen:

- nicht erkannten Zeichen → werden von OCR-Software i. d. R. entsprechend markiert
- falsch erkannten Zeichen → müssen im konvertierten Text mühsam gesucht werden

Serifenlose Zeichensätze sind solche, bei denen die Zeichensätze keine Endstriche an Zeichen haben, z. B. Arial, Helvetica oder Noto Sans (dieser Foliensatz verwendet Noto Sans (Display)).

Schriftarten mit Serifen sind z. B. Times New Roman oder Garamond.

Barcodes und QR-Codes

- Werden zur Identifizierung von Dokumenten eingesetzt.
- 2 Einsatzgebiete:
 1. Selbst erzeugte Dokumente (z. B. Anträge) mit Barcode-Aufdruck: Beim Rücklauf automatisch erkennbar.
 2. Für Fremddokumente: Barcode-Etiketten (Szenario: „Spätes Archivieren“).
- Sehr robust und etabliert.
- Bar-/QR-Codes weisen sehr hohe Erkennungsraten auf.

Beispiel

Lohnsteuerbescheinigung mit QR-Code

Ausdruck der elektronischen Lohnsteuerbescheinigung für 2023

Nachstehende Daten wurden maschinell an die Finanzverwaltung übertragen.



Szenarien: Zeitpunkt des Scannens

Drei typische Erfassungsszenarien für Eingangspost:

- Scannen im Posteingang (frühes Archivieren)
- Scannen zum Zeitpunkt der Bearbeitung
- Scannen nach der Bearbeitung (spätes Archivieren)

Szenario 1: Frühes Archivieren / Erfassen beim Posteingang

- Eingehende Dokumente werden vor der eigentlichen Bearbeitung gescannt
 - Scannen erfolgt meist im Posteingang
 - Weiterleitung an Sachbearbeiter auf elektronischem Weg
- Vor elektronischer Weiterleitung: evlt. Klassifikation + evtl. Attributierung

Vorteil: Elektronische Weiterleitung

- ✓ Kurze Transportzeiten,
geringe Transportkosten
- ✓ Weiterleitung an mehrere Personen
- ✓ Evlt. automatisierte Adressermittlung
- ✓ Steuerung und Verfolgen der
Bearbeitung (Workflow)

Nachteile:

- ! Sachbearbeiter benötigen Arbeitsplatz
mit DMS-Zugang
- ! ggf. Neuausrichtung des Geschäftsprozesses
- ! ggf. aufwändiger Einstieg

Szenario 2: Erfassung bei der Bearbeitung

- Dokumente gelangen in Papierform zum Sachbearbeiter.
- Dort werden sie direkt vor oder gleich nach der Bearbeitung eingescannt, attributiert und abgelegt.

Einsatzgebiet

- Erfassung, Nachbearbeitung oder Attributierung ist aufwendig oder erfordert spezielle Sachkenntnis
- fehlgeleitete Belege werden in das DMS eingebracht
(Ggf. in Ergänzung zum „Frühen Archivieren“.)
- kleine Dokumentenmengen, nicht für Massenbearbeitung geeignet

Nachteile

- ! Bearbeitungsplätze müssen mit Scanner ausgestattet sein
- ! ständiger Wechsel zw. Dokumentenerfassung und Bearbeitung stört Arbeitsfluss
- ! Einsatz teurer Personalressourcen (Sachbearbeiter) für einfache Tätigkeiten
(Scannen, Attributieren)

Szenario 3: Spätes Archivieren

- Papierdokumente werden nach ihrer Bearbeitung an die zentrale Erfassungsstelle geschickt und dort eingescannt.
- Zusätzlich wird ein Identifikator für das Papierdokument benötigt.
 - für Zuordnung des Papierdokuments zu Vorgang während Bearbeitung
 - Bar-/QR-Code oder Referenznummer/Belegnummer
- Bar-/QR-Code:
 - Registrierung: Dokument erhält eindeutigen Barcode z. B. im Posteingang oder durch Sachbearbeiter
 - Barcode-Erfassung mit Barcodestift oder Lesepistole
 - Erfassung des Papierdokuments
 - Erfassungssoftware erkennt Code automatisch
 - Code auf der ersten Seite kann gleichzeitig für Dokumententrennung genutzt werden
 - Die Zuordnungstabelle zw. Code und Dokument ist regelmäßig zu prüfen, ob alle registrierten Dokumente zwischenzeitlich gescannt wurden

Typischerweise werden die Codes nach Erfassung des Dokuments nicht mehr benötigt und eine Wiederverwendung ist ca. nach einem Jahr möglich.

Szenario 3: Spätes Archivieren - Bewertung

Vorteile

- ✓ Arbeits- und Papierflüsse können weitgehend wie bisher abgewickelt werden.
- ✓ Papierdokumente (z. B. Rechnungen) können vor ihrer Erfassung noch geprüft und abgezeichnet werden: Stempel, Unterschrift, Korrekturen werden beim Scannen erfasst.
- ✓ Arbeitsplätze der Sachbearbeiter erfordern keine spezielle Ausstattung.

Nachteile

- ! Eigentliches Potenzial elektronischer Dokumente wird nicht genutzt.
- ! Gefahr des Verlusts oder der Beschädigung des Papierdokumentes höher.

Scannen von Dokumenten - Zusammenfassung

Entscheidungsdimensionen:

- frühes Scannen vs. spätes Scannen oder Scannen bei der Sachbearbeitung
- zentrales Scannen vs. dezentrales Scannen
- scannen und indizieren gleichzeitig oder zeitlich versetzt
- Selbst scannen oder Outsourcing (externer Dienstleister)

3. COLD-Verfahren

(ursprünglich Computer Output on Laser Disk)

COLD

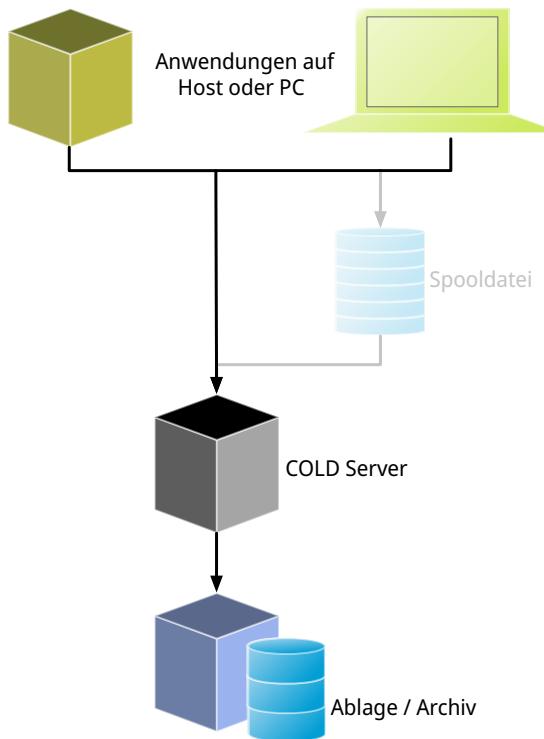
Begriff stammt aus der Zeit Mitte der 80er Jahre, hatte sich aber bereits zu Beginn/Mitte der 90er technologieunabhängig verallgemeinert.

Beschreibt **die direkte digitale Speicherung von Druck- und Listenausgaben betrieblicher Softwaresysteme** (z. B. direkt von ERP Systemen oder von Office Anwendungen über spezielle Druckertreiber).

- Die Recherche kann danach wie bei jedem anderen Dokument im DMS erfolgen.
 - COLD bei größeren Unternehmen bzw. DMS-Lösungen sehr verbreitet.
 - COLD-Verarbeitung ist typische Batch-Verarbeitung.
-

d. h. bei COLD werden die Daten nicht mehr - bzw. nur optional - auf Papier ausgegeben, sondern stattdessen direkt in ein DMS übernommen. Da kein OCR notwendig ist, sondern die Daten direkt „beim Drucken“ abgegriffen werden, ist die Qualität der Daten sehr hoch.

COLD-Verfahren (historisch)



Verarbeitung auf COLD-Server

1. Zerlegung des Datenstroms in einzelne Dokumente.
2. Extrahiert die für die Ablage bzw. spätere Recherche der Dokumente notwendigen Index-Daten automatisch + evtl. Bezug zu Overlays.
(Die Fachdaten und das Layout sind getrennt.)
3. Konvertierung bringt die Dokumente in eine für die Ablage geeignete Form.

4. Metadaten für Dokumente

Metadaten

- Beschreibende Merkmale für Dokumente
- Ziel ist das möglichst exakte Wiederfinden der richtigen Dokumente (strukturierte Suche!)
- Metadaten sind strukturiert und möglichst exakt vordefiniert (z. B. Wertebereiche)
- Quellen für Metadaten:
 - Manuelles Erfassen
 - Aus dem Dokument automatisch ermitteln
 - Aus anderen Anwendungen / Quellen übernehmen

Manuelles Indizieren

- Freitexteingabe (z. B. Zusammenfassung, Notizen)
- Unterstützung durch Auswahlmenüs, Formatvorgaben oder Defaultwerte, z.B.
 - Schlagwortindizierung (definierter Wortschatz)
 - Formalisierte Eingabe (z. B. Datum)

■ **Probleme:**

- ! Fehleranfällig
- ! Aufwändig
- ! Ergebnis vom Bearbeiter abhängig

(Semi-)Automatisches Indizieren

■ basierend auf wissensbasierten bzw. regelbasierten Ansätzen

Durch ein umfangreiches Regelwerk wird versucht, die Metadaten (insbesondere Art des Dokuments, Vorgangsnummer, Empfänger) automatisch zu ermitteln; um eine automatische Klassifikation und Verarbeitung zu ermöglichen.

■ basierend auf (überwachten) maschinellen Lernverfahren

Das System wird in einem ersten Schritt - basierend auf eingescannten Dokumenten - überwacht trainiert und kann dann in einem zweiten Schritt die Metadaten automatisch ermitteln.

Suche und Retrieval von Dokumenten

Strukturierte Suche

Unter Nutzung der Metadaten werden gezielte Anfragen an das DMS gestellt.

- ✓ Suche per Daten über Dokumente, die nicht unbedingt direkt in den Dokumenten zu finden sind.

! Suchraster ist vorgegeben (d. h. Metadatenschema ist fest).

Volltextsuche

Wenn die Dokumente als CI-Dateien vorliegen, dann kann man auch mittels Volltext suchen. Evtl. ergänzt um semantische Hilfsmittel (Thesaurus, etc.).

- ✓ Man kann jedes Wort wiederfinden.

! Unstrukturiert, „langsam“, Ressourcenbedarf, keine semantisch zusammenfassenden Informationen abfragbar.