Dokumentenerfassung und -indizierung

Dozent: **Prof. Dr. Michael Eichberg**
Kontakt: michael.eichberg@dhbw-mannheim.de, Raum 149B
Version: 2024-02-16

Dieser Foliensatz basiert auf Folien von: Klaus Götzer.

Alle Fehler sind meine eigenen.

Dokumenten-Management von *Klaus Götzer, Patrick Maué, und Ulrich Emmert*, dpunkt.verlag, 2023.

DHBW
Duale Hochschule
Baden-Württemberg
Mannheim

1. Quellen von Dokumenten

Prof. Dr. Michael Eichberg

Quellen von Dokumenten - Dimensionen

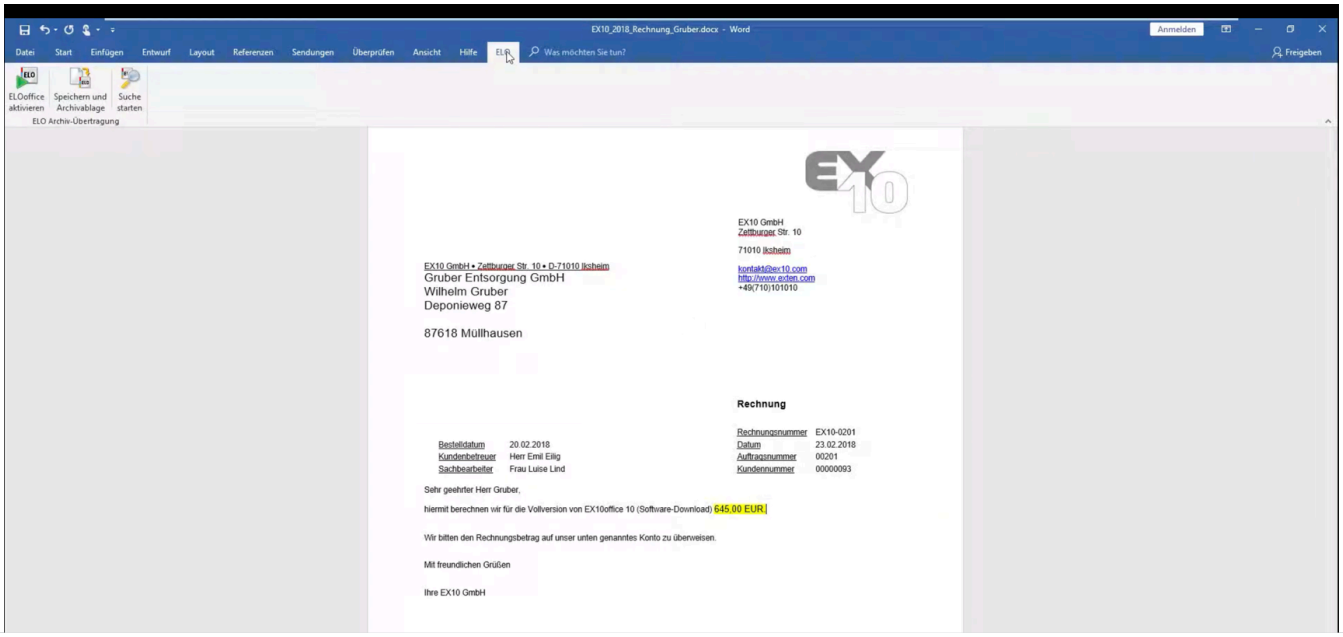
- Eigenerstellte und fremderstellte Dokumente
- Papierdokumente und elektronische Dokumente
- Einmalige Übernahme und laufende Übernahme

Eigenerstellte Dokumente

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

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

Speichern von Dokumenten aus Anwendungen



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

Elektronische Dokumente

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

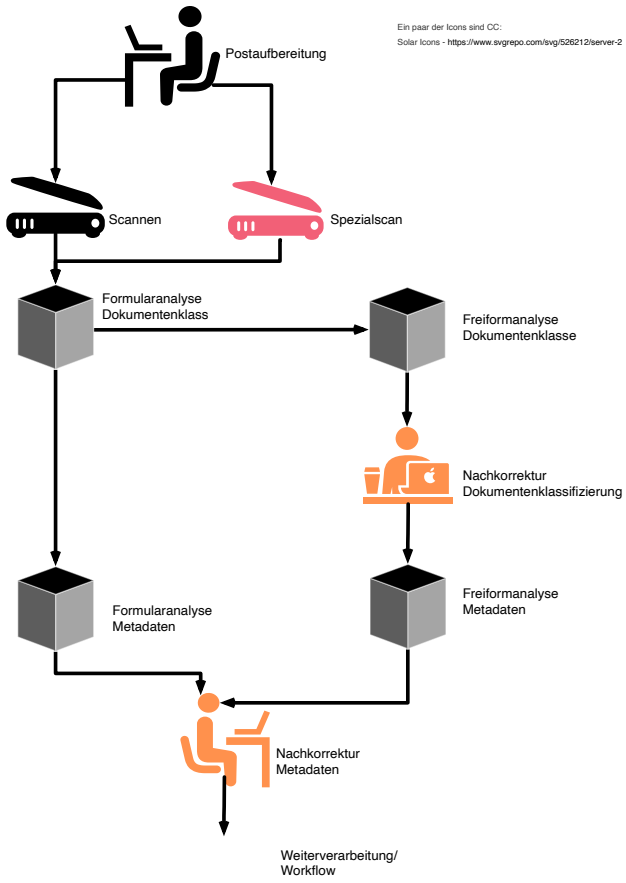
  Strukturiertes Dokument oder Fließtext?

Papierdokument

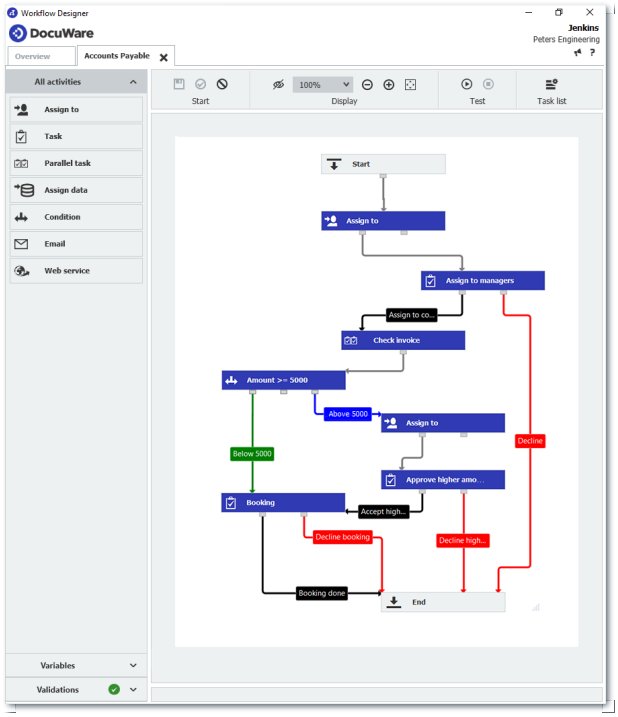
- S/W oder farbig?
- Automatisch auszuwerten?
- Aufwand für manuelle Vorbereitung (Entheften, Glätten, ..)

Beispiel für Eingangspostbearbeitung

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



Unterstützung für Workflowdefinitionen in ECM Systemen



https://start.docuware.com

ECM: *Enterprise Content Management*

Erstmalige Übernahme von Dokumenten

Quellen

- Altsystem (Archiv, DMS)
- Filesystem
- Mikrofilm, Mikrofish 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.

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

Prof. Dr. Michael Eichberg

Scannen der Eingangspost

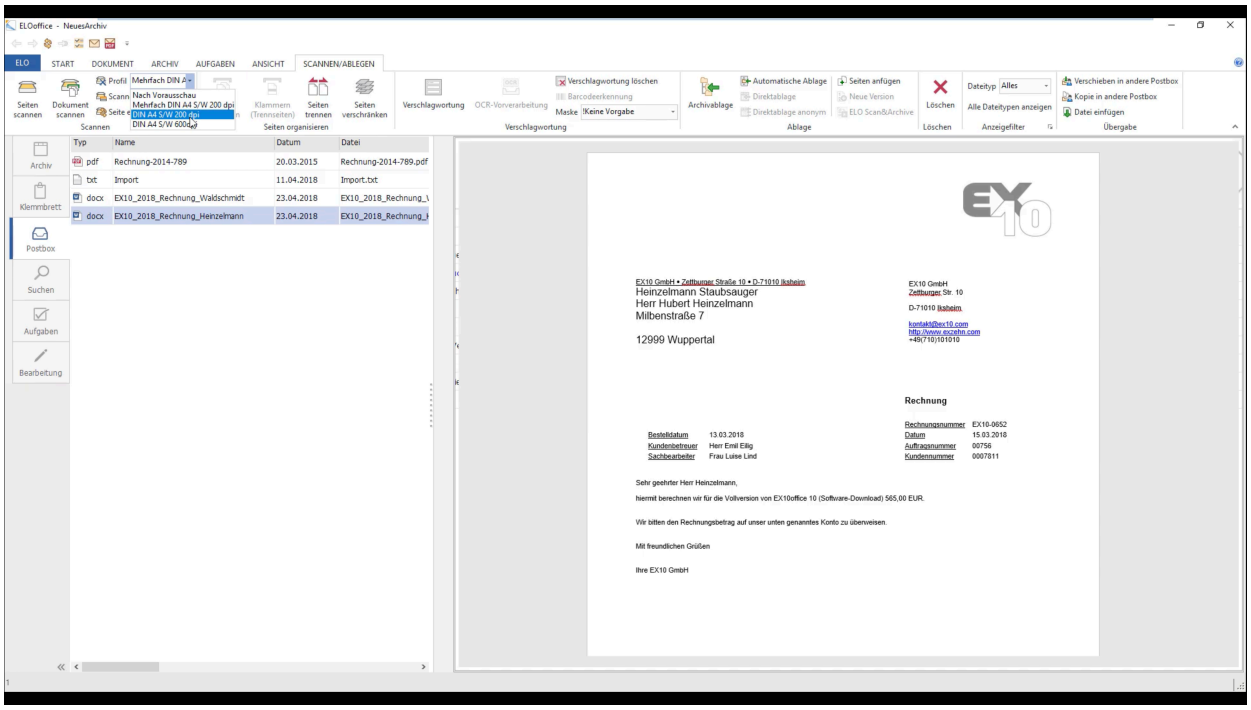
- Scanner: meist verbreitetes Erfassungsgerät für Dokumente auf Papier oder Film

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)



1

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 Dokumentenklasse bekannt ist)

Umwandlung von NCI zu CI

Optical Charakter Recognition (OCR):

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

Handprint Charakter Recognition (HCR):

Erkennen von handschriftlichen Texten.

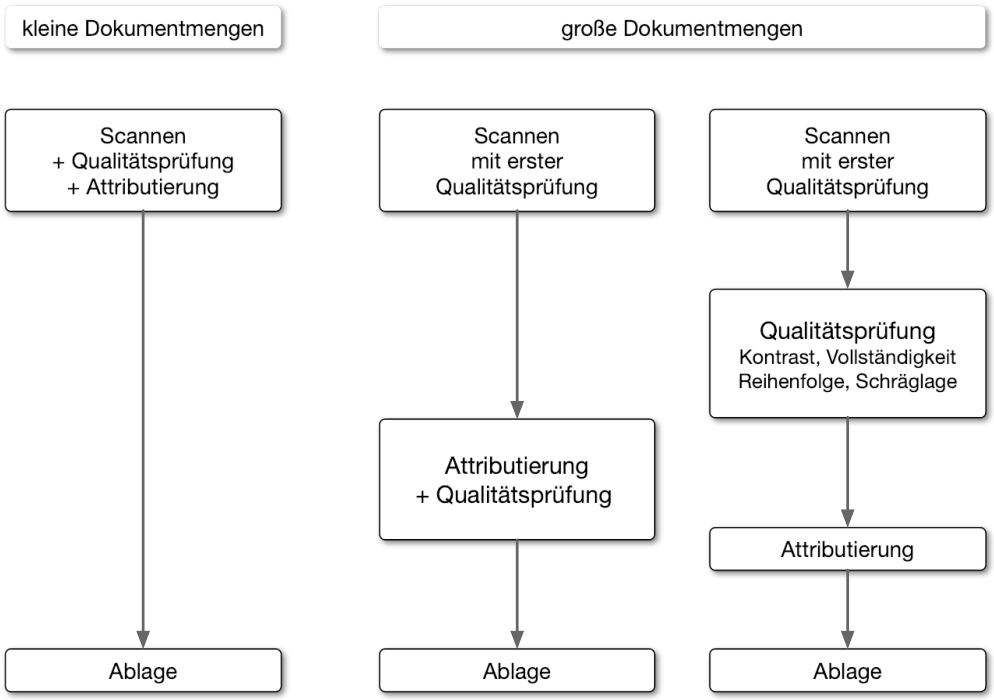
Intelligent Charakter Recognition (ICR):

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 (wie z.B. Selektionsfelder aus Fragebögen oder 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...)

1

Problemfälle

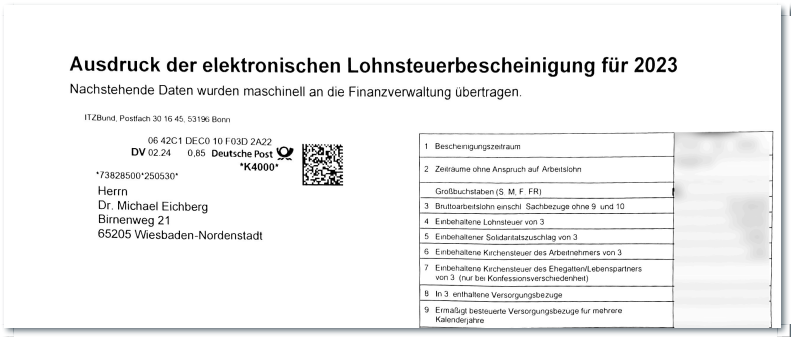
- Ligaturen (z.B. ffi `statt ffi oder fi statt fi`)
- Bestimmte Zeichenkombinationen z.B. rn: ‚r' gefolgt von ‚n' oder ‚m'
- Großes I (wie Ida) und kleines l (wie lieb) bei serifenlosen Zeichensätzen
- Fremdsprachige Zeichen (z.B. $)

20

Serifenlose Zeichensätze sind solche, bei denen die Zeichensätze keine Endstriche an Zeichen haben. z.B. Arial oder Helvetica.

Barcode/ QR-Code

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



Ausdruck der elektronischen Lohnsteuerbescheinigung für 2023

Nachstehende Daten wurden maschinell an die Finanzverwaltung übertragen.

ITZBund, Postfach 30 16 45, 53196 Bonn

06 42C1 DEC0 10 F03D 2A22
DV 02.24 0,85 Deutsche Post
K4000

*73828500*250530*

Herrn
Dr. Michael Eichberg
Birnenweg 21
65205 Wiesbaden-Nordenstadt

1 Bescheinigungszeitraum
2 Zeiträume ohne Anspruch auf Arbeitslohn
Großbuchstaben (S, M, F, FR)
3 Bruttoarbeitslohn einschl. Sachbezüge ohne 9. und 10.
4 Einbehaltene Lohnsteuer von 3
5 Einbehaltener Solidaritätszuschlag von 3
6 Einbehaltene Kirchensteuer des Arbeitnehmers von 3
7 Einbehaltene Kirchensteuer des Ehegatten/Lebenspartners von 3 (nur bei Konfessionsverschiedenheit)
8 In 3. enthaltene Versorgungsbezüge
9 Ermäßigt besteuerte Versorgungsbezüge für mehrere Kalenderjahre

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: Erfassen beim Posteingang (*Frühes Archivieren*)

- 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)

Nachteil:

- ! 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 Dokumentenbearbeitung stört Arbeitsfluss
- ! Einsatz teurer Personalressourcen (Sachbearbeiter) für einfache Tätigkeiten (Scannen, Attributieren)

Szenario 3: Spätes Archivieren

- Papierdokumente werden nach ihrer Bearbeitung an 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.
  - Code wird nach Erfassung des Dokuments nicht mehr benötigt; Wiederverwendung ist ca. nach 1 Jahr

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 elekt. Dokumente wird nicht genutzt
- ! Gefahr des Verlusts oder der Beschädigung des Papierdoks höher

Zusammenfassung

- 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 (Computer Output on Laser Disk)

Prof. Dr. Michael Eichberg

COLD

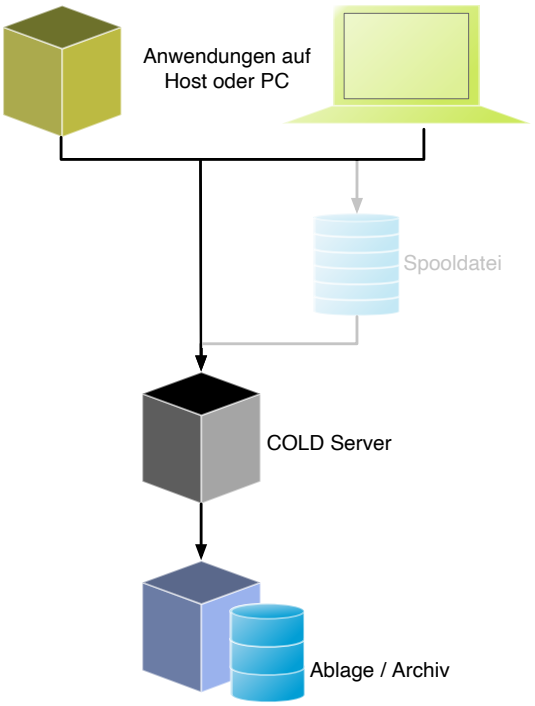
Begriff stammt aus der Zeit Mitte der 80er Jahre, hatte sich aber bereits zu Beginn/Mitte der 90er technologieunabhängig verallgemeinert.

Beschreibt **die direkte digitale Speicherung von 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 COLD-Server

1. Zerlegung des Datenstrom in einzelne Dokumente
2. Extrahiert die für die Ablage bzw. spätere Recherche der Dokumente notwendigen Index-Daten automatisch + evtl. Bezug zu Overlays (Trennung zwischen fachlichen und layout Daten)
3. Konvertierung bringt die Dokumente in eine für die Ablage geeignete Form

DHBW
Duale Hochschule
Baden-Württemberg
Mannheim

4. Metadaten für Dokumente

Prof. Dr. Michael Eichberg

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

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.).

- ✓ Vorteil: Man kann jedes Wort wiederfinden.
- ! Unstrukturiert, „langsam”, Ressourcenbedarf, keine semantisch zusammenfassenden Informationen abfragbar