

**E-DAY 2003**

# **OCR- und Scannerintegration im Dokumentenmanagement**

**Petra Hauschke - RWS WISSENSMANAGEMENT GmbH, Balingen  
Reinhard Merz – RM Informationstechnologie GmbH, Dotternhausen**

**Wenn Ihr Unternehmen  
wüsste, was es alles weiß...**

**Thomas H. Davenport 1998**

# Elektronisches Archiv – warum?

- **Optimierungsmöglichkeit der Prozesse**
- **Material- und Raumkosten sparen**
- **Zeitersparnis – finden nicht nur suchen**  
(Annahme: pro Mitarbeiter 30 min Suchen/Tag → 3 Wochen Suchzeit/Jahr  
→ jährlich ca. 5.000 Eur/pro Mitarbeiter)
- **Risikominderung**
- **Informationsflut bewältigen**

# Kostenvergleich

Datenträger	Kapazität in Seiten	Anzahl Ordner je 400 Seiten	Datenträgerkosten	Ordnerkosten	Platzbedarf
CD	6.500	16	0,50 EUR	32,00 EUR	1,3 Meter
DVD	52.000	130	5,00 EUR	260,00 EUR	10 Meter
Festplatte 30 GB	300.000	750	150,00 EUR	1.500,00 EUR	60 Meter

Quelle: [www.aktien-james.de](http://www.aktien-james.de) (Stand 2002)

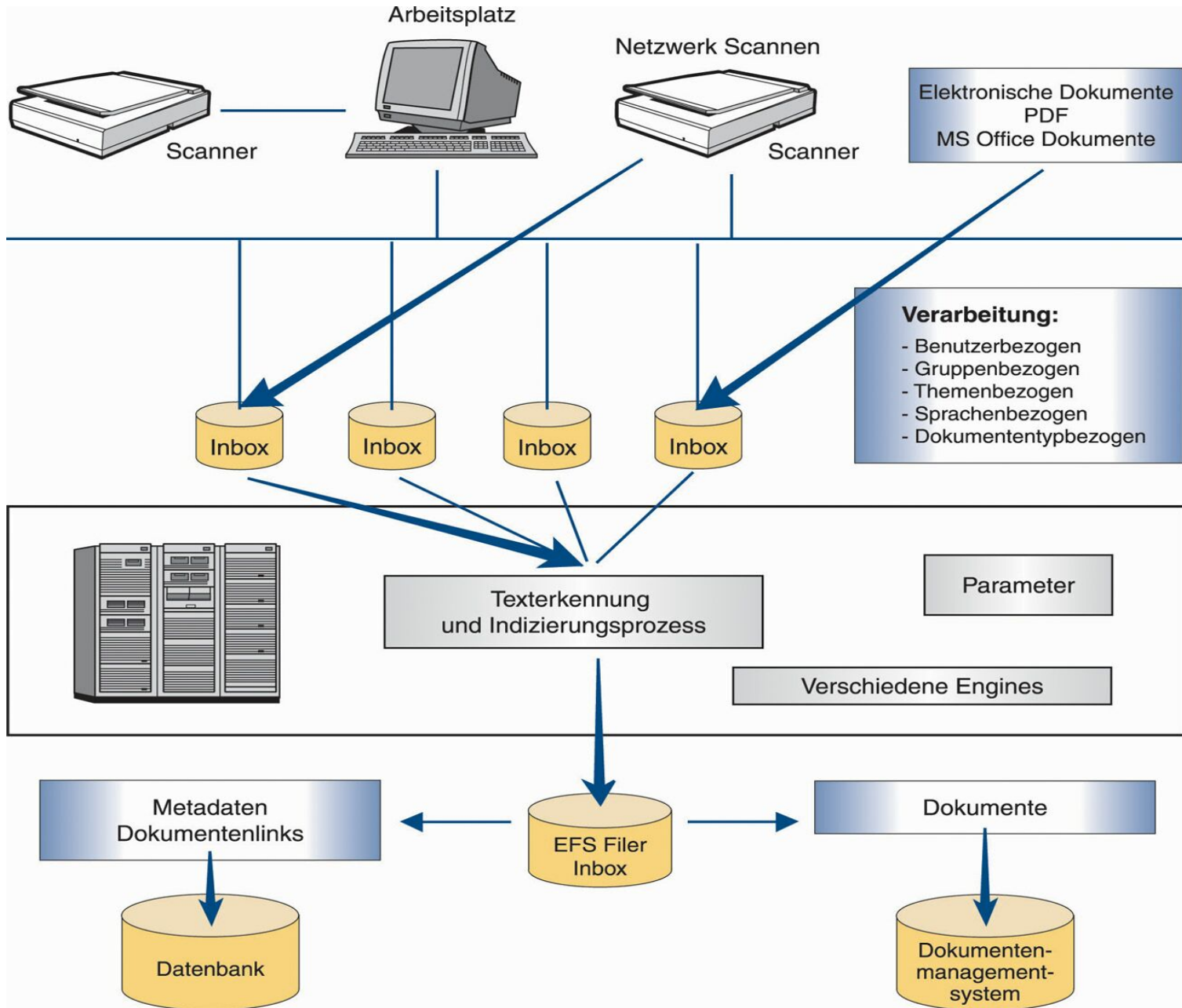
# Scanner < - > OCR

Scanner → Bilddatei (Format : multipage Tiff)



Bilddatei → OCR (Optical Character Recognition = Texterkennung)

# Prozessablauf



# Mira Suite

---

- **Mira EFSOCR**
- **Mira EFSIndex**
- **Mira EFSFiling**
- **Mira EFSLoad**

# Mira EFSOCR

---

- **Automatischer Hintergrundprozess**
- **Extrahiert aus Bild-, PDF-, Microsoft Office- Dateien Volltext**
- **Berücksichtigt vorgegebenen Parameter**  
(Benutzerbezogen, Gruppenbezogen,Themenbezogen,Dokumentbezogen)
- **Ergebnis → XML - Datei**

# Mögliche Vorgaben

---

- **Spracheinstellungen (122 Sprachen)**

Unterscheidungsmöglichkeit neue - alte deutsche Rechtschreibung (Bsp.: Delphin – Delfin)

- **Benutzer-, GruppenID, FirmenID**

- **PDF-Erstellung**

- **Permanente Stichworte**

# Mira EFSINDEX

---






























- **Automatischer Hintergrundprozess**
- **Erstellt aus EFSOCR – Ergebnis -  
basierend auf vorgegebenen Parameter-  
eine semantischen Analyse**
- **Ergebnis → XML – Datei**  
(mit Volltext, kontrollierten Stichwörtern und am häufigsten  
vorkommenden Begriffe)

# Mögliche Vorgaben

---

- **Stichwortliste**
- **Hierarchische Liste (Thesaurus)**
- **Stopwortliste**
- **Bearbeitungsverzeichnisse**

# Hierarchischer Thesaurus

-  [Medientyp](#)
-  [Organisation](#)
-  [Rechtsbereich](#)
-  [Steuerberatung](#)
-  [Unternehmensberatung](#)
  -  [Unternehmensführung und Organisation](#)
    -  [Organisation](#)
      -  [Informatik](#)
      -  [Informatik Anwendungen](#)
      -  [Informationstechnik und Informatik](#)
      -  [Informationstechnik und Systemarchitektur](#)
      -  [Programmierung und Software](#)
        -  [Betriebssystem](#)
        -  [Programmiersprache](#)
          -  [C++](#)
          -  [Cobol](#)
          -  [DHTML](#)
          -  [HTML](#)
          -  [Java](#)
          -  [Pascal](#)
          -  [Perl](#)
          -  [Shell Script](#)
          -  [Visual Basic](#)
          -  [XML](#)
        -  [Programmierung](#)
        -  [Software](#)
        -  [Systemanalyse](#)
-  [Betriebliche Information und Kommunikation](#)
-  [Rechtsformen](#)
-  [Unternehmensentwicklung, Betriebsgröße und Standortwahl](#)
-  [Umweltmanagement](#)

# Mira EFSFiler & Loader

---

- **Automatischer Hintergrundprozess**
- **Speichert die Dateien EDMS**
- **Meta-Daten und Dokumentlinks werden in die Datenbank geladen**

# Prozessablauf

