Strategien der Langzeitspeicherung von Digitalisaten bei den Staatlichen Archiven Bayerns

Überblick

- Grundprämisse
- Nutzung
- Workflow für Digitalisierung
- Metadaten
- Bildformat
- Erhalt der Digitalisate
- Digitalisierungskonzept

Grundprämisse

Zweck der Digitalisierung

- Sicherheitsdigitalisierung
  - Notfallkopie im Falle des Verlustes
- Schutzdigitalisierung
  - Erleichterung der Nutzung
  - Verbreitung des Nutzerkreises
  - Erschließen neuer Nutzungsmöglichkeiten
  - Schutz des Original gegen weitere Abnutzung
- Ersatzdigitalisierung
  - Erhalt der Information anstatt des Originals

Grundprämisse

- Ersatzdigitalisierungen haben den Rang von Archivalien

→ Erhalt ist oberste Prämisse (Digitales Archiv)

- Digitalisate sind keine Archivalien

→ Wirtschaftlichkeit ist bestimmend

- Digitalisierung selbst ist nur Teil der Kosten
- Kostentreiber sind Erhaltung und Nutzung

→ Strategie wird bestimmt durch Erhaltung und Nutzung

Nutzung

Nutzung der Digitalisierung

- Verschiedene Nutzungen sind möglich
  - Internetpräsentation
  - Lesesaalkopien (Ersatzvorlagen)
  - Vorlagen zur Öffentlichkeitsarbeit
  - Benutzerkopien
    - zum Lesen
    - zur Veröffentlichung
    - zur Nachbearbeitung
    - ...

Anforderung der Nutzungen

- Jede Nutzung hat eigene Anforderungen
  - schnell zu öffnen
  - geringer Speicherverbrauch
  - verbreitete Lesesoftware
  - detailreich
  - scharf
  - ...

→ Spezielle Formate für Nutzung notwendig

Anforderungen für die Langzeitspeicherung aus Nutzung

- Jedes gewünschte Nutzformat muss sich aus dem Langzeitformat erstellen lassen
- Langzeitformat kann unterschiedlich zu den Nutzformaten sein
- Höchste Qualitätsanforderung der Nutzung bestimmt Anforderung an Langzeitkopie
- Digitalisat muss gefunden werden können
- Digitalisat muss nachbearbeitbar sein
- Rückgriff muss kostengünstig sein
- Rückgriff muss hinreichend schnell genug sein

Verwendung festlegen

- Nutzung bestimmt Qualität der Digitalisate
  - Höchste mögliche Auflösung ist nicht hilfreich
  - **Qualität dery Scans ist entscheidend**
- Größe bestimmt den Preis der Speicherung
- Spekulation über zukünftige Nutzung ist kaum möglich

→ Verlustbehaftete Komprimierung ist möglich, wenn sie die zu erwartende Nutzung erlaubt

→ Jetzige Nutzung bestimmt die Digitalisierung

Workflow für Digitalisierung

Digitalisierung in der Organisation

- Digitalisierungsprojekte betreffen die ganze Organisation
- Digitalisierung allein hat keinen Nutzwert

→ Zu Beginn eines Projektes müssen

→alle Ressourcen gesichert sein

→alle Phasen geplant sein

→ Auch für Nutzung und Langzeiterhaltung

Massenproduktion von Digitalisaten

→ Standardisierung und Workflow

- Kostensenkung durch Standardisierung
- Fehlerreduktion durch Automatisierung
- Findbuch als führende Applikation

→ Verknüpfung der Digitalisate mit dem Findbuch notwendig

Notwendige Ressourcen

- Verzeichnung ist Vorrausetzung für Digitalisierung
- Personal und Gelder müssen bereitgestellt werden
- URN für Präsentation

Anforderungen des Workflows an die Langzeitspeicherung

- Langzeitspeicherung muss sich in den Workflow einfügen
  - Workflow für Digitalisierung
  - Workflow für Nutzung
- Langzeitspeicherung muss massenfähig sein

Digitalisierung im Rahmen von Benützeraufträgen

- Einzelproduktionen
- teilweise teure Digitalisate (Karten, Urkunden...)
- teilweise sehr empfindliche Stücke
- Einzelverwaltung ist unwirtschaftlich

→ Integration in Workflow

→ rudimentäre elektronische Verzeichnung ist notwendig

Ziele der Langzeitspeicherung

- Primäre Ziele
  - Schutz der Investition in Digitalisierung
  - Ermöglichen der Nutzung über einen lange Zeit
    - EDV Zeiträume sind 5 Jahre
- Sekundäre Ziele
  - Schutz der Originale

Bildformat

Bildformat für Langzeitspeicherung

- Auswahl aus wenigen anerkannten Formaten (TIFF, PNG, JPG 2000)
- Nestor empfiehlt TIFF oder JPG 2000
- Prämisse: Speicherplatz ist teuer und bestimmt Erhaltungspreis
- Format muss offen sein und Migration in ein neues Format erlauben

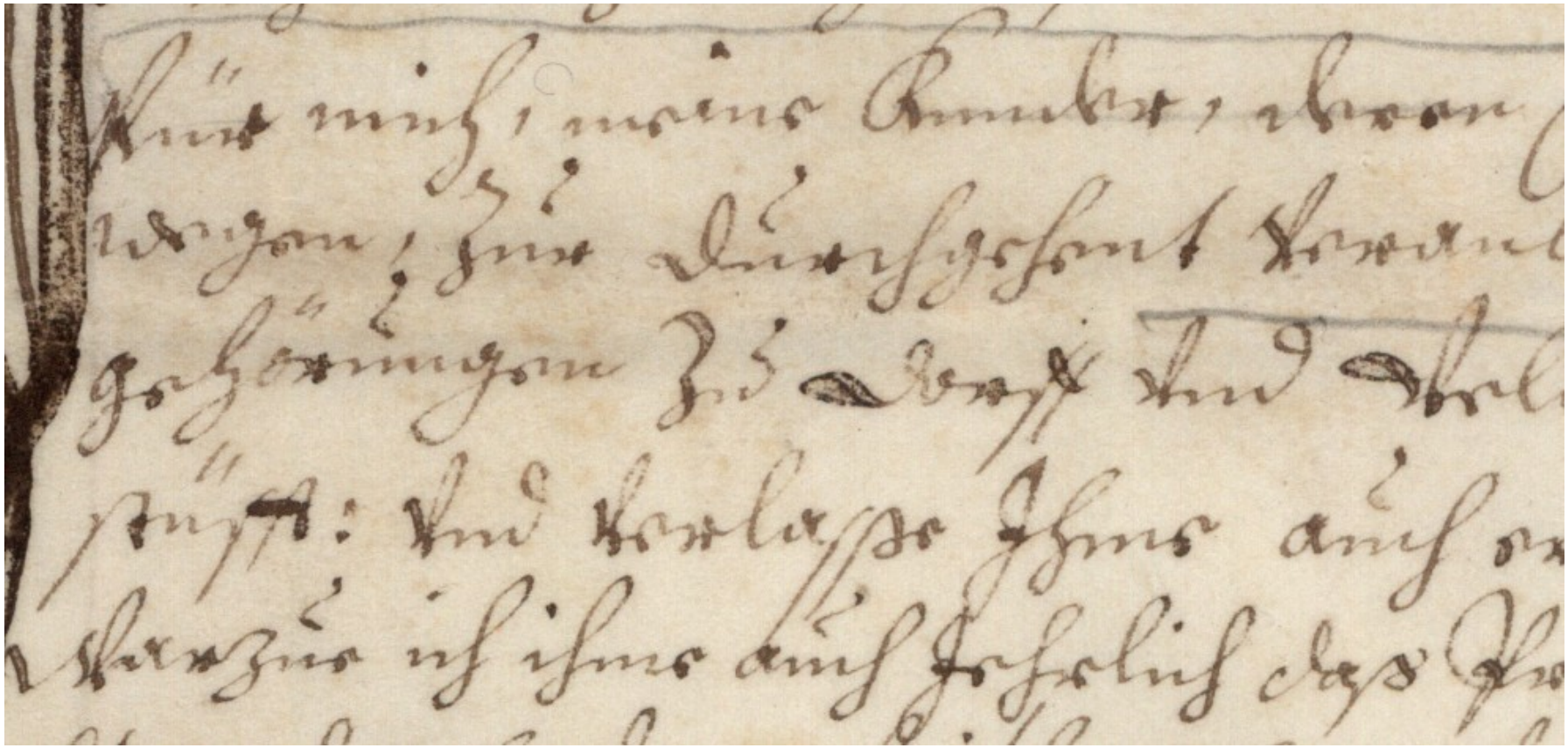
Kompression der Bildformate

- Kompression stark abhängig vom Ausgangsmaterial
- PNG komprimiert für Zeichnungen gut
- JPEG 2000 spielt seine Stärke bei Bildern aus
- Bei TIFF sind LZW und ZIP üblich
- Hohe Qualität des Digitalisats führt zu besseren Kompressionsergebnissen

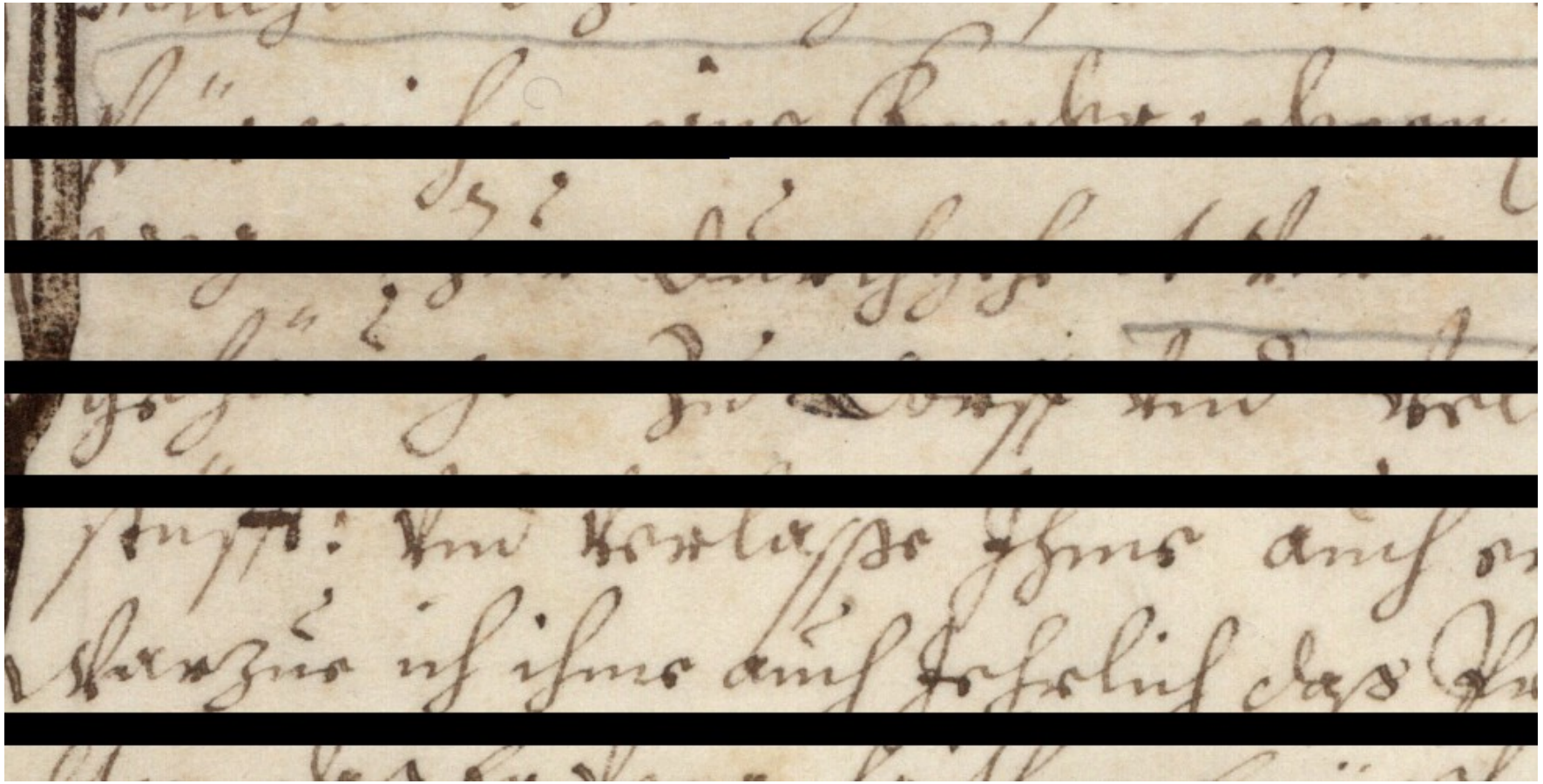
Fehler im Langzeitspeicher

- Bei der Definition von JPEG 2000 wurde Wert auf eine Toleranz gegen Übertragungsfehler gelegt.
  Siehe auch Annhang A8, D5, J4 sowie K6 des Standards (ISO/IEC 15444-1)
- Am wahrscheinlichsten ist der Verlust von ganzen Speicherblöcken
- Simulation des Verlusts von 1024 Blöcken á 512 Bytes (Original ~ 160 MByte)

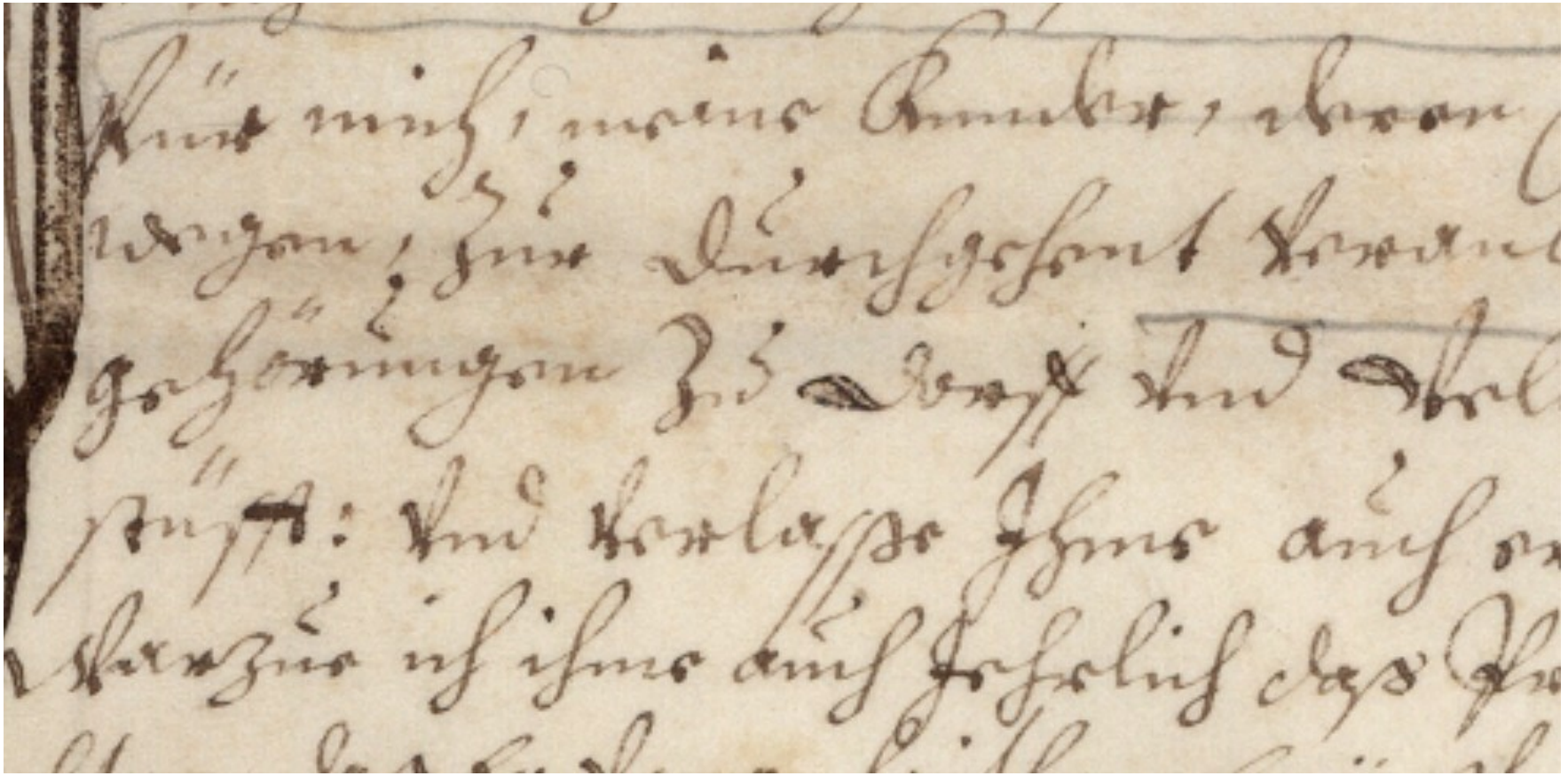
Original Bild



Fehler bei TIFF



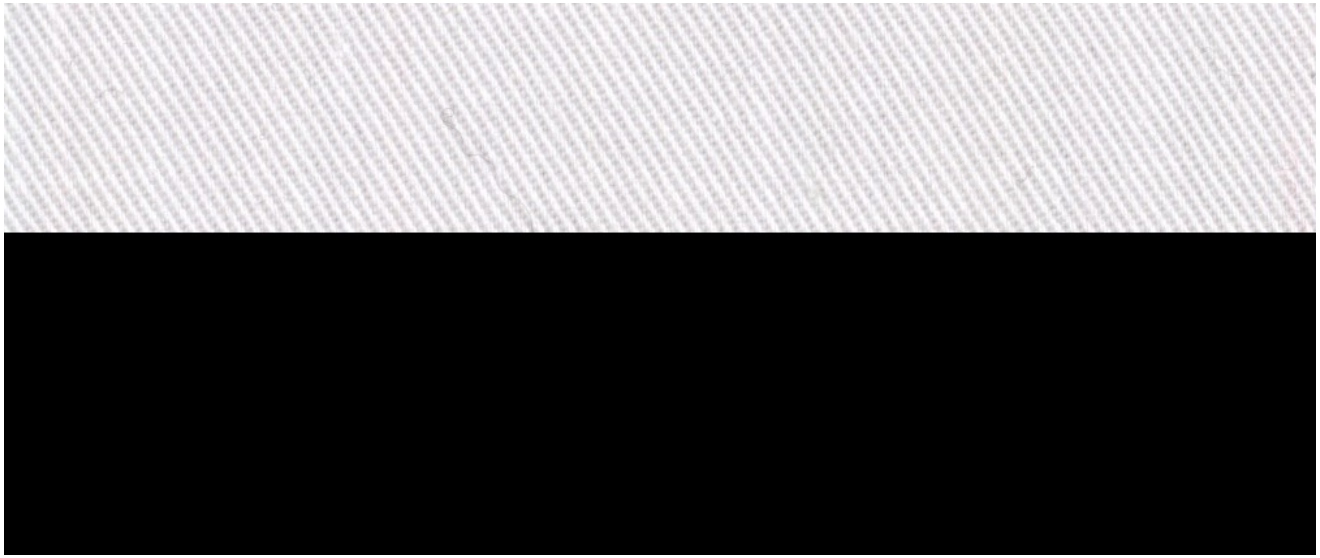
Fehler bei JPEG 2000



Fehler im Bildformat

- Simulation der Fehler in komprimierten Formaten unter gleichen Bedingungen
  - PNG verliert gesamte Information
  - komprimiertes TIFF verliert die gesamte Information
  - komprimiertes JPEG 2000 verhält sich wie unkomprimiertes JPEG 2000 - es wird unschärfer

PNG / komprimiertes TIFF



Offenheit

- *Part 1 der JPEG 2000 Spezifikation*
  - *definiert den Kern von JPEG 2000*
  - *JPEG 2000 codestream, encoding, decoding*
  - *Basic File Format*
  - *Basic Metadaten wie Farbraum*
  - *Internationaler Standard (ISO/IEC 15444-1)*
  - *weitere Teile des Standards sind nicht essentiell*
- *Basis JPEG 2000 sollte lizenzfrei sein*
  - *Eine formale Garantie gibt es dafür nicht*
- *quelloffene Referenzimplementierung Jasper existiert*

Verwendetes Langzeitformat

JPEG2000 als stark komprimierendes Format

- Unempfindlichkeit gegenüber Fehlern
- verlustfreie und verlustbehaftete Kompression
- Base Level ist offen und frei

→ Nur Base Level

Metadaten

Metadaten

- Wert der Digitalisate bleibt nur mit Metadaten erhalten
- Nicht alle Metadaten sind im Findbuch

→ Metadaten zu den Digitalisaten

- Standardisierung der Metadaten
  - Automatisierung der Verarbeitung
  - Erleichterung der Nutzungsprozesse

Klassifikation Metadaten

- archivfachliche Metadaten
  - Daten des Findbuchs
    - (Bestand, Bestellnummer ...)
- technische Metadaten
  - Metadaten des Bildes
    - (Auflösung, Größe, Farbprofil ....)
  - Entstehungsgeschichte und Verarbeitung
    - (Scanstation und Umgebungsbedingungen, Verarbeitungsschritte)

Speichern der Metadaten

- Version A im Bild
  - \+ einfache Verwaltung nur einer Datei
  - – Extraktion nur mit speziellen Zusatztools
  - – Migration in neues Format erschwert
  - – Einschränkung bei der Formatwahl
- Version B in getrennter XML-Datei
  - \+ freie Migration von Bild und Metadaten
  - \+ Verwendung von Standard-Tools
  - – getrennte Verwaltung von zwei Dateien

Erzeugen der Metadaten

Erzeugen der Metadaten

- Tools zur automatisierten Extraktion der Metadaten
- Verknüpfung Metadaten / Findbuch mit ID
- automatisierte Extraktion der technischen Metadaten
- Erstellung der Bilddatei aus dem bearbeiteten Scan
- Kombination aller Daten zum Digitalisat

Erhalt der Digitalisate

Erhaltung der Digitalisate

- zentrales NAS am Standort München
  - örtliche Nähe zu größten Erzeugern und Nutzern
- NAS in den Staatsarchiven
  - örtliche Nähe zu deren Beständen und Nutzern
- Sicherung
  - dezentrale NAS auf zentrales NAS
  - zentrales NAS auf Bänder
  - wichtige Bestände auf NAS im Rechenzentrum
  - NAS im RZ auf Band

Einspeichern der Digitalisate

- Digitalisat besteht aus Bilddaten und Metadaten
- Extraktion ID (UUID) und Pfad aus Metadaten und Speichern in Abbildungsdatenbank
- Extraktion der Prüfsumme in Prüfsummendatenbank
- Einspeichern der Metadaten und der Bilddaten in die Bilddatenspeicher gemäß den Metadaten

Einspeichern der Digitalisate

Problem der Verfälschung

- Degeneration
  - ECC Mechanismen begrenzt wirksam
  - Speicher verfällt mit der Zeit
- Übertragungsfehler
  - Änderung während Umkopieren und DFÜ
- Geringe Zugriffsfrequenz

  → übliche Mechanismen der Rechner HW greifen nicht

  → Fehler werden zu spät erkannt → Verlust

Lösungsmöglichkeiten

- Lösungen aus dem Enterprise Bereich
  + sehr zuverlässig
  + wartungsarm
  - sehr teuer
- Einfache Lösungen
  – wartungsaufwendiger
  + problemadäquate Lösung
  + günstiger

Sicherstellung gegen Verfälschung

Sicherstellung gegen Verfälschung

- jeder Speicher wird regelmäßig auf Veränderung geprüft
  - Berechnung der Prüfsummen für alle in der Abbildungsdatenbank gehaltenen Digitalisate
  - Vergleichen der errechneten Prüfsummen mit den Prüfsummen der Prüfsummendatenbank
  - Erstellung eines Fehlerprotokolls
  - Periode durch Backup bestimmt

Verknüpfung von Digitalisat und Findbuch

- Digitalisat muss gefunden werden
- Findbuch ist führende Applikation
- technisches System
  → Kopplung durch technische ID
- UUID als dezentral zu erstellender Schlüssel zur Verknüpfung

Wiederauffinden des Digitalisats

- DB zur Verknüpfung von Speicher und Findbuch
- Trennung von Speicherort und Applikation

→ Verschiedene Ausprägungen der Digitalisate möglich
(z.B. Online / Lesesaal / Benutzerkopie)

→ Migration von Bildspeicher und Applikation unabhängig möglich

Wiederauffinden des Digitalisats

Tools zur Unterstützung der Langzeitspeicherung

- Extraktion der Metadaten aus dem Findbuch
- Erzeugen eines Digitalisat nach der Bearbeitung
- Einspeicherung eines Digitalisat
- Rücklesen eines Digitalisat
- Checken der Digitalisate
- Mapping-Applikation

Digitalisierungskonzept

Offizielles Digitalisierungskonzept

- Digitalisierungskonzept wurde bis Mitte 2012 ausgearbeitet
- Absegnung durch GDA und Verteilen an nachgeordnete Archivbehörden
- wichtig, damit dieses Konzept umgesetzt wird
- Aufnahme aller Ausnahmen ins Konzept
- Abweichungen vom Konzept sind verboten

Inhalte des Konzeptes

- Workflowbeschreibung
- Festlegung des Bildformates
- Festlegung der Verwendung und der Auflösungen für die einzelnen Archivalientypen
- Festlegung der Metadaten
- Festlegen der Tools zum Workflow

Einführung

- Implementierung erfolgt schrittweise
  1. Einführung im Bayerischen Hauptstaatsarchiv und zentraler Fotowerkstatt
  2. Einführung in den Staatsarchiven
- Kontinuierliche Verbesserung durch

FRAGEN?

Michael.Kirstein@gda.bayern.de