

Open Digital Information/Image Archive (OpenDIA) Workflow-Konzept

Oliver Flimm <flimm@openbib.org>
Stand: 28. Mai 2005

Inhaltsverzeichnis

1 Aufgabe	3
2 Grundkonzept	4
2.1 Anforderungen	4
2.1.1 Flexibilität	4
2.1.2 Einfachheit in der Bedienung und Administration	4
2.1.3 Kosten	4
2.2 Konzept	4
2.2.1 Scan-Vorgang	4
2.2.2 Anreicherung mit Meta-Daten	5
2.2.3 Präsentation	8
2.2.4 Integration	8
2.2.5 Fazit	8

1 Aufgabe

Die zu bewältigende Aufgabe besteht in der Erstellung, Beschreibung, Darstellung und Verbreitung von Digitalisat-Serien. Bei den Digitalisaten handelt es sich primär um Folgen von digital eingescannten Bildern, die eine logische Einheit bilden - z.B. ein Buch, einen Einband, ein Akzessionsjournal usw. In sich können diese Digitalisate eine hierarchische Struktur mit verschiedenen Ordnungseinheiten aufweisen, wie z.B. Kapitel, Sektionen, Untersektionen usw.

Mit diesem Papier soll ein Konzept vorgestellt werden, wie ein Einfachstansatz für einen Workflow realisiert werden kann, mit dem diese Aufgaben zu bewältigen sind.

Grundsätzlich sind in dem Workflow folgende Schritte abzudecken:

Scan-Vorgang In diesem Schritt werden die Bild-Dateien in einer größtmöglichen – aber dennoch sinnvollen – Auflösung erzeugt. Diese Daten müssen gespeichert werden.

Anreicherung mit Meta-Daten Zusammengehörige Bild-Dateien bilden ein logisches Digitalisat und müssen mit Meta-Daten angereichert werden. Diese können das gesamte Digitalisat, einzelne Ordnungseinheiten (z.B. Kapitel usw. mit festgelegter Abfolge) oder einzelne Bilder beschreiben und müssen u.a. auch die Gesamt-Struktur der Bilder bzgl. Ordnungseinheiten und Abfolgen beschreiben (entsprechend Inhaltsverzeichnissen/Ebenen).

Damit ist der Hauptteil der zu leistenden Arbeit abgedeckt. Die Anreicherung mit Meta-Daten bindet das meiste Personal und stellt den aufwändigsten Teil dar. Nachdem alle Meta-Daten erfasst sind, folgen diese Schritte:

Präsentation Die Bilder müssen mit ihren Meta-Daten und ihrer Ordnungsstruktur im Web präsentiert werden. Das stellt ein geringes Problem dar, da sich prinzipiell programmtechnisch alle denkbaren Browsing/Darstellungsarten anhand der Metadaten und der Bilder erzeugen lassen. Denkbar sind sowohl teils statische Webseiten, die im Voraus erzeugt werden müssen, oder dynamisch just-in-time generierte Webseiten, die durch ein geeignetes Programm angezeigt werden.

Integration Durch geeignete Mechanismen muss eine Integration in andere Dienste – unabhängig von einer angestrebten eigenständigen Nutzbarkeit – möglich sein. Hierzu gehört die lokale Integration in Nachweisinstrumente wie den KUG, aber auch die Integration in externe Systeme, z.B. via OAI oder WebServices.

2 Grundkonzept

2.1 Anforderungen

2.1.1 Flexibilität

Die zu bewältigende Aufgabe ist bezogen auf ihre wesentlichen Schritte sehr klar strukturiert. Dennoch kann für jeden dieser Schritte der Teufel im Detail stecken. Erfahrungsgemäß werden erst nach und nach konkrete Anforderungen (z.B. speziell das zu verwendende Meta-Datenformat) formuliert, so daß davon auszugehen ist, dass die erste Version in der Realisierung nicht die letzte bleibt. Daher ist eine maximale Flexibilität der anvisierten Lösung unabdingbar.

2.1.2 Einfachheit in der Bedienung und Administration

Die Lösung muß letztlich durch das vorhandene Personal administriert, angepasst, erweitert und bedient werden. Aus diesem Grund sollte sie

- auf bereits Bekanntem aufsetzen und
- keine neuen, nicht beherrschten Technologien einsetzen.

2.1.3 Kosten

Ebenso spielen die Kosten eine wesentliche Rolle. Kommerzielle Lösungen sind in der Regel nicht sehr preiswert. Durch die Verwendung von bereits vorhandener bekannter Software und bekannten Technologien sowie von OpenSource-Produkten können die Kosten minimiert werden.

2.2 Konzept

Unter Berücksichtigung der Anforderungen kann ein Workflow wie folgt realisiert werden.

2.2.1 Scan-Vorgang

Die Dateien werden vom zuständigen Personal gescannt und in einem Verzeichnisbaum auf einem Netz-Laufwerk *strukturiert und darin in der jeweils logischen Abfolge* entsprechend vorher getroffener Vereinbarungen gespeichert. Strukturiert bedeutet eine Verzeichnisstruktur auf dem Laufwerk in der Form

```
.../collection/digitalisat/kapitel/unterkapitel/etc.
```

Die generelle Organisation erfolgt in Digitalisat-Serien (Collections), in denen Digitalisate abgelegt werden.

Durch die Verwendung von Verzeichnissen unterhalb `digitalisat` lassen sich hierarchische Strukturen des Dokuments allein mit den Mittel eines Dateisystems abbilden. So könnten schon beim Scan die Verzeichnisse, Unterverzeichnisse etc. auf dem Laufwerk erzeugt werden, die dann einem Kapitel, Unterkapitel etc. entsprechen. Wenn keine (Unter)kapitel etc. existieren, dann vereinfacht sich die Organisation der Bild-Daten entsprechend.

Beispiel ohne Kapitel für die Digitalisat-Serie 'einbaende':

```
.../einbaende/1/0001.tif
.../einbaende/1/0002.tif
.../einbaende/1/0003.tif
.../einbaende/1/0004.tif
.../einbaende/1/0005.tif
```

Beispiel mit Kapitel für die Digitalisat-Serie 'einbaende':

```
.../einbaende/1/0001.tif
.../einbaende/1/0002.tif
.../einbaende/1/kap_1/0003.tif
.../einbaende/1/kap_1/0004.tif
```

Bei Digitalisaten aus katalogisierten Büchern entspräche `digitalisat` sinnvollerweise dem numerischen Katalogschlüssel. In den Beispielen wäre der entsprechende Katalogschlüssel in der zugehörigen Katalog-Datenbank, in den Beispielen z.B. 1. Über eine solche Vereinbarung wäre eine Datenübernahme aus den Katalogdaten problemlos möglich.

Die zu verwendenden Werkzeuge dieses Schrittes sind bereits vorhanden und in der Bedienung bekannt. Es ist der Scanner samt Software und ein Datei-Browser (Windows Explorer oder KDE Konqueror).

Als Ergebnis dieses Schrittes liegen die Bild-Daten in der korrekten Reihenfolge und Strukturierung auf dem Netz-Laufwerk vor.

2.2.2 Anreicherung mit Meta-Daten

Für das Anlegen von Meta-Daten (und zusätzliche Arbeiten für die Web-Präsentation, die auch hier angesiedelt/angestossen werden könn(t)en) sind verschiedene Modelle denkbar. Ein solches Modell soll nun vorgestellt werden.

Meta-Daten-Mapping über das Datei-System

Der Grundgedanke ist: *Meta-Informationen werden in Dateien mit geeigneten Namen an geeigneter Stelle abgespeichert, so dass eine Zuordnung zu einer Digitalisat-Serie, einem Gesamtdigi-*

talizat, (Unter)Kapitel oder Bild möglich ist.

Das bedeutet in der konkreten Realisierung:

- Alle Meta-Daten werden in Dateien abgelegt, die eine spezifische Endung (z.B. `.dsc`) besitzen.
- Pro Kategorie eines Meta-Datums zu einem Bezugs-Objekt (Digitalisat-Serie, einzelnes Digitalisat, einzelnes Bild einzelnes Kapitel etc.) wird eine Datei verwendet.
- Jede Meta-Datenkategorie wird im Meta-Daten-Dateinamen kodiert.
- Die Dateien von Meta-Daten zu einer hierarchischen Ebene (Digitalisat-Serie, einzelnes Digitalisat, Kapitel etc.) beginnen mit `meta`, an welches getrennt durch ein definiertes Zeichen (z.B. `_`) der Name der Meta-Datenkategorie angehängt wird.
- Die Dateien von Meta-Daten zu einer (Bild-)Datei beginnen mit dem zugrunde liegenden (Bild-)Dateinamen, an den getrennt durch ein definiertes Zeichen (z.B. `_`) der Name der Meta-Datenkategorie angehängt wird.
- Multiple Inhalte einer Meta-Datenkategorie werden in der entsprechenden Datei voneinander getrennt (z.B. durch `' - '`)

Beispiel: Willkürliche Meta-Information 'Beschreibung' via Meta-Datenkategorie `DC.Subject`

Beschreibung des Gesamtdokuments/Digitalisates:

```
.../einbaende/1/meta_DC.Subject.dsc
```

Beschreibung eines Bildes `0001.tif`:

```
.../einbaende/1/0001.tif_DC.Subject.dsc
```

Beschreibung des Kapitels 1:

```
.../einbaende/1/kap_1/meta_DC.Subject.dsc
```

Beschreibung des Bildes `0003.tif` in Kapitel 1:

```
.../einbaende/1/kap_1/0003.tif_DC.Subject.dsc
```

Ein großer Vorteil dieses Ansatzes ist

- die Möglichkeit der sofortigen Verwendung beliebiger Meta-Datenkategorien, ohne das zugrunde liegende Datenmodell jeweils dem neuen Meta-Datenmodell anpassen zu müssen,
- die Möglichkeit auch nachträglich – nach der Vergabe von Meta-Daten (s.u.) – die Struktur des Dokuments durch einfache Operationen des Datei-Browsers auf dem Netzwerklaufwerk zu ändern.

Web-basierter Editor für die Meta-Datenverarbeitung

Der Web-basierte Editor greift grundsätzlich auf den Festplattenbereich zu, auf dem im vorangegangenen Schritt die Scans abgelegt wurde. Damit kann die Web-Schnittstelle von sich aus schon auf die so strukturierten Bild-Dateien zugreifen.

Der Benutzer bekommt in einer Auswahl alle Digitalisat-Serienverzeichnisse

```
.../collection/...
```

auf dem Netz-Laufwerk angeboten, von der er die Digitalisat-Serie (collection) auswählt, die er bearbeiten möchte.

Daraufhin bekommt er die vorhandenen einzelnen Digitalisate

```
.../.../digitalisat/...
```

angezeigt, von denen er sich eines auswählen kann. Unabhängig davon kann der Benutzer an dieser Stelle Meta-Daten für die Digitalisat-Serie vergeben.

Daraufhin bekommt er eine Darstellung der Dokumentbestandteile/struktur angezeigt.

An dieser Stelle kann der Benutzer die entsprechenden Meta-Daten für das Digitalisat als Ganzes, einzelne Bestandteile oder Ordnungseinheiten eingeben.

Zusätzlich könnten an dieser Stelle verschiedene Schritte angesiedelt sein, die für die Web-Veröffentlichung relevant sind und von hier aus anstossen werden könnten. Dazu könnte die automatische Erzeugung von Pictogrammen/Thumbnails der gescannten Bilder sowie eine oder mehrere spezielle Versionen für die Web-Darstellung gehören.

Hier wird auch die Möglichkeit der Datenübernahme aus einem bereits bestehenden Katalog angeboten werden.

Die Abspeicherung der hier eingegebenen Metadaten erfolgt in dem oben dargelegten Text-Dateischema. Eine direkte Speicherung in einer XML-Struktur wäre ebenso möglich. Gleiches gilt für den Aufruf eines schon bearbeiteten Digitalisats. Die vorher aufgenommenen Meta-Informationen können aus dem Text-Dateischema extrahiert werden, oder aus einer vorher abgespeicherten XML-Datei.

Ebenso können hier automatisch URN's generiert, an die DDB gemeldet werden usw. Hier ist fast alles automatisierbare ansiedelbar.

Dieser Ansatz bietet folgende Vorteile:

- Einfache Bedienung
- Einfache Eingabemöglichkeit auch vieler (gleichartiger) Meta-Informationen 'auf einen Rutsch'
- Kopplung mit dem Katalogsystem problemlos möglich zur automatischen Meta-Datenübernahme bzw. -erzeugung.
- Maximale Flexibilität

2.2.3 Präsentation

Ausgehend von den Metadaten und den Bildern wird ein Programm erstellt, um die Digitalisate zu präsentieren. Hier ist die Integration einer MySQL-Datenbank sinnvoll, in die nach Fertigstellung eines Digitalisats oder einer Digitalisierungs-Serie die entsprechenden Informationen – Meta-Daten und Scan-Daten – abgelegt werden können, um diese dann bei der Präsentation (z.B. beim Browsing) schnell verwenden zu können. Generell ist dieser Teil – wenn die Meta-Daten einmal existieren – sehr einfach zu realisieren.

2.2.4 Integration

Durch die Verwendung der MySQL-Datenbank sind aus den dort angesiedelten Informationen alle denkbaren Ausgabeformate generierbar. Diese können XML-basiert sein und z.B. via OAI angeboten werden. Ebenso ist durch die Offenheit der Anwendung eine Integration in andere Systeme problemlos möglich.

2.2.5 Fazit

Der dargestellte Einfachstansatz eines Workflow-Konzeptes auf Grundlage des verwendeten Mappings in das Dateisystem

Scannen -> Meta-Daten Editor -> Präsentation/Integration

ist beliebig modifizierbar, soll aber im Wesentlichen zeigen, dass alle Phasen des Workflows auch mit sehr einfachen Bordmitteln (Scanner samt Program, Datei-Browser und Web-Browser) und etwas Programmierung bewältigt werden können.

Der Ansatz ist so offen, daß beliebige Meta-Datenformate (auch parallel) für beliebige Digitalisierungs-Serien realisiert werden können. Jedes Meta-Datenformat muß lediglich in einer Konfigurationsdatei mit seiner Struktur abgelegt werden. Eine Änderung der Programme, des Datenmodells usw. ist nicht notwendig.

Dieser Ansatz realisiert damit genau die an die Lösung gestellten Anforderungen:

- Simpel und daher auch einfach zu administrieren, erweitern und bedienen.
- Maximale Flexibilität (Anpass- u. Erweiterbarkeit)
- Minimale Kosten