

# Erläuterungen zur Digitalisierung altertümlicher Bücher, ihrer Erschließung mit Hilfe der Software Goobi, nebst den Möglichkeiten ihrer Präsentation

## Prinzipielles

Ein Scan ist eine Reproduktion einer physischen Bildvorlage in Form eines digitalisierten Bildes. Eine Buchseite kann grundsätzlich als ein Bild aufgefaßt werden. Folglich handelt es sich beim Vorgang des Scannens um eine Abbildung eines physischen Seitenbildes des Buches auf ein digitales Bildformat. Grundsätzlich geht dabei keinerlei Information verloren, legt man eine ausreichende Qualität der Reproduktionstechnik zugrunde, d.h. solange technische Verluste ausgeschlossen werden können, enthält das Abbild – und zwar prinzipiell – exakt dieselben Informationen wie das Urbild. Was allerdings aufgelöst wird, ist die physische Einheit Buch, in dem sämtliche Einzelseiten der inhaltlichen Einheit des geistigen und schriftlichen Werkes Buch zusammengefaßt sind. Da pro Scan einer Buchseite eine einzelne Bilddatei erstellt wird, ist das Buch zunächst fragmentiert. Rudimentär wiederhergestellt wird die ursprüngliche Einheit durch das Abspeichern aller Einzelbilder in einem entsprechend benannten Dateiordner.

Ein Buch ist nun in mehrfacher Hinsicht strukturiert: Zum einen ist da die rein physische Struktur, bestehend aus den einzelnen Blättern, Einband, etc. Wichtiger ist dann die formale Struktur – Titelei, Verzeichnisse, Anhänge, etc. Wesentlich jedoch ist die inhaltliche Struktur: angefangen bei der Untergliederung in Kapitel bis hinunter zu Wörtern, Punkt und Komma und am Ende zu Sätzen und Worten.

Wie aus dem ersten Absatz ersichtlich ist, wird die physische Struktur in der Abbildung (nur) teilweise erhalten, ist aber auch perfekt rekonstruierbar: Aus einem Blatt des Buches werden zwei Seiten(-bilder); die Unterteilung nach Seiten eines Buches ist allerdings bereits Teil der formalen Struktur, so daß zumindest diese Struktur perfekt abgebildet wird.

Was weder in den *Urbildern* noch in den *Abbildern* unmittelbar vorhanden ist, sind die formale – abgesehen von der Paginierung – und vor allem die inhaltliche Struktur. Diese wird im wesentlichen mental vom Leser rekonstruiert. Grundlage hierfür sind die *Bildinhalte*, nicht die Bilder selbst.

Die gedanklichen Prozesse, die beim Lesen eines Buches ablaufen, sind von erstaunlicher und zugleich atemberaubender Komplexität. Ein Computer ist ver-

gleichsweise verblüffend und sagenhaft dumm. Die inhaltliche Auswertung eines Buches liegt Lichtjahre jenseits dessen Fähigkeiten. Ein Computer wird ein Buch niemals *verstehen*. Gewisse bescheidene Annäherungen sind aber mittlerweile möglich. Andererseits übersteigt die Schnelligkeit von Computern, in den Feldern, auf denen die Rechner Fähigkeiten besitzen, die menschliche bei weitem. Eine statistische Auswertung etwa ist für Menschen beinahe eine Lebensaufgabe. Hier sollen zunächst zwei Dinge festgehalten werden. Erstens eine inhaltliche Rekonstruktion von Büchern analog dem menschlichen Lesen ist Computern mit den digitalisierten Seitenbildern nicht möglich. Eine inhaltliche Erschließung von digitalisierten Büchern bietet Lesern dagegen ein großes Potential in Bereichen, in denen der Mensch durch seine Langsamkeit gehandicapt ist.

Aufgrund der bislang gemachten Überlegungen stellen sich nun mehrere Fragen.

1. Wie kann aus der Ansammlung digitalisierter Bilder ein Produkt erstellt werden, das dem ursprünglichen Werk soweit nahekommt, daß es für den Leser eine hinreichende Ähnlichkeit mit dem Ausgangswerk erhält, soll heißen: eine vergleichbare Nutzung ermöglicht?
2. Wie kann die erwähnte inhaltliche Erschließung erfolgen? Und welchem Zweck genau soll sie dienen?
3. In welcher Qualität, in welchem Format und in welcher Größe sollen die digitalisierten Bilder vorgehalten werden?

Einen guten Einstieg bietet Frage zwei: Computer unterscheiden fundamental zwischen Bild und Text – solange wir nicht auf der binären Ebene nachsehen. Ein erster Schritt ist daher, aus Bildern wieder Text zu machen, das sogenannte OCR-Lesen. Der so ermittelte Text wird den Ausgangsbildern hinterlegt. Auf diese Weise wird sowohl eine Volltextsuche ermöglicht, als auch das Herauskopieren von Text aus den Digitalisaten. Die Bilder bleiben als mehr oder weniger exaktes Faksimile der Buchseiten erhalten. Können, müssen aber nicht – je nach Wunsch des Auftraggebers.

Ein zweiter Schritt kann die formal-inhaltliche Erschließung des Buches sein. Selbstverständlich beginnend mit der Zusammenfassung aller Einzelbilder zu dem Gesamtwerk Buch. (In welcher Form ist hier die Anschlußfrage als Antwort auf die Frage eins oben.) Abgesehen davon geht es um die Strukturierung der Ansammlung von Bildern innerhalb der Einheit Buch. Relevant sind hierfür folgende Aspekte: Zunächst die Reihenfolge der Bilder nebst der Paginierung der Seiten. Hierzu erhält jedes Bild zwei Attribute, neben einer fortlaufenden Nummer für die Abfolge der Bilder, die schlicht durch eine Numerierung im Dateinamen erfolgen kann, wird jedem Bild eine Seitenzahl zugeordnet als zusätzliche Informationseinheit.

Auf einer nächsthöheren Ebene der Strukturierung geht es um die Wiederherstellung – und das heißt um die Generierung einer maschinenverwertbaren Form – der Gliederung des Buches. Systematisch basiert das ganze auf folgendem Ansatz: Es werden Strukturelemente gebildet, die sich aus zwei Komponenten zusammensetzen. Der Kennzeichnung des Elements zum einen und der Zuordnung von Seitenbildern zum anderen. Außerdem können Strukturelemente typisiert werden. Als Beispiel nehmen wir ein Strukturelement vom Typ Kapitel, gekennzeichnet durch den Überschriftentext des Kapitels, dem eine bestimmte Reihe von Bilddateien zugeordnet ist, deren Reihenfolge sich dann aus deren Numerierung, sowie der zugewiesenen Paginierung ergibt. Daraus sollte das Prinzip deutlich werden.

Ermöglicht wird so eine Übersicht über das Werk, ohne sich jede einzelne Seite anschauen zu müssen. Gleichzeitig kann über eine Verlinkung der Strukturelemente dem Nutzer ermöglicht werden, per Mausklick an die entsprechende Stelle im Buch zu springen.

Schlußendlich kann auf einer übergeordneten Ebene noch die bibliographische Erschließung des Werkes erfolgen. Wiederum technisch schlicht und einfach durch die Addition einer Informationseinheit zum Gesamtbestand des Digitalisats, die sämtliche, soweit gewünscht, bibliographischen Daten zu einem Buch enthält. Eine solche Informationseinheit gestattet dann die edv-mäßige Suche nach und Verarbeitung der Bücher als Gesamtheiten.

Aus all diesen Bausteinen kann eine einzelne Datei zusammengesetzt werden, der die Bilder samt dem zugehörigen Volltext beigegeben werden. Es gibt also immer noch eine Vielzahl von einzelnen Dateien, wobei sich jedoch in einer Weiterverarbeitung durch Nutzung der Informations- und Strukturdatei das Buch wiederherstellen läßt. Nennen wir das die Präsentationsvariante.

Eine zweite Möglichkeit ist die Zusammenfassung aller zugehörigen Informationseinheiten – inklusive des Volltextes auf den Buchseiten – und Bilder zu einer einzigen Datei. Diese wird von einem geeigneten Programm geöffnet und zeigt die Seiten des Buches nebst aller verfügbaren Informationen in einheitlicher und zusammengehöriger Form, die ein Buch näherungsweise repräsentieren kann. Üblicherweise nutzt man für solche Zwecke das PDF-Dateiformat.

Goobi ist in der Lage beide solche Dateien zu erstellen. Es bietet Werkzeuge, alle benötigten Informationen zu erstellen, die einem Buch, sprich den Seitenbilddigitalisaten, zugeordnet werden sollen. Diese Informationen werden in den benötigten Dateiformaten ausgegeben.

Selbstverständlich immer erhalten bleiben die faksimilierten Seitenbilder als Resultat des Scanvorgangs. Sie sind grundsätzlich unabhängig von allen folgenden Verarbeitungsschritten und weiteren Informationseinheiten je für sich zugänglich und werden gewissermaßen als Originale gesichert.

## Die Weiterverarbeitung

PDF-Dokumente sind plattformunabhängig nutzbar. Es handelt sich um einzelne Dateien pro Buch. Sie sind daher sehr gut handhabbar, die Angelegenheit bleibt auch in der Archivierung sehr übersichtlich, außerdem bietet sich die Chance auf einfache, simple Downloads, die ein Anbieter zur Verfügung stellen kann.

Schwachpunkt ist die praktische Unmöglichkeit der Weiterbearbeitung einer solchen PDF-Datei. Bilder sind nur mit erheblichem Aufwand wieder zu entnehmen. Informationen nur sehr eingeschränkt zu ändern. Die einmal gewählte Qualitätsstufe der Bilder, wie auch die Bilder selbst nicht mehr zu korrigieren.

In der Präsentationsvariante kehren sich Vorzüge und Nachteile der PDF-Ausgabe praktisch um. Es müssen sämtliche einzelnen Bilder auf Servern vorgehalten werden, auf die zum Aufruf des Buches aktuell ein Zugriff möglich sein muß. Es bleibt eine Vielzahl von Einzeldateien vorhanden. Es wird ein Programm zur Präsentation benötigt, das den in den Strukturdaten verwendeten Katalog von Typen an Strukturelementen verarbeiten kann. Der Zugriff erfolgt üblicherweise über ein Netzwerk, so daß die Verarbeitungsgeschwindigkeit abhängig ist von der – meist niedrigen – Netzwerk- oder Internetgeschwindigkeit.

Andererseits bleiben alle Bilder als separate Dateien vorhanden und sind beliebig zu bearbeiten oder zu konvertieren. Die Strukturdaten können jederzeit beliebig und hochflexibel geändert oder korrigiert werden. Das ganze findet überdies zentral statt, so daß nicht im Anschluß weiterhin fehlerhafte Dateien unterwegs sind, wie das etwa bei bereits verbreiteten PDF-Dokumenten der Fall wäre.

Der Aufwand, die Bilddateien vorzuhalten, ist allerdings erheblich höher, als das bei PDFs der Fall ist. Überdies muß ein Präsentationsprogramm online verfügbar gemacht werden. Das andererseits wiederum den Vorzug der wesentlich höheren Flexibilität in der Gestaltung der Präsentation gegenüber der PDF-Ansicht bietet.

Ein vorhandenes und frei zugängliches Präsentationsprogramm ist der öffentlich verfügbare – d.h. online nutzbare DFG-Viewer. Es wird auf der eigenen Website ein Link auf diesen DFG-Viewer gesetzt, der den Pfad auf die mit Goobi erstellte Informationsdatei (im Dateiformat .xml) enthält, die auf dem eigenen Server liegt. Der DFGViewer greift auf diese XML-Datei zu und entnimmt ihr die Pfade auf die Bilddateien des Buches, die ebenfalls auf dem eigenen Server liegen. Auf diese Weise kann jeder externe Nutzer über den DFG-Viewer auf Bücher zur Ansicht zugreifen, die auf dem anbietereigenen Server bereitgestellt werden, ohne daß hierzu eine weder benutzereigene, noch anbieterseitig eingerichtete Präsentationsstufe oder ein entsprechendes Programm vorhanden sein müßte. Schnittstelle wäre der via Internet frei zugängliche DFG-Viewer.

## Die Dateiformate

Die von uns gelieferten Dateien gliedern sich in vier Gruppen.

Die erste Gruppe bilden die PDF-Dateien. Zu diesen sollten keine weiteren Erläuterungen nötig sein, sie liefern in sich abgeschlossene Endprodukte, die von jedem ohne weitere mitzuliefernde Hilfsmittel genutzt werden können.

Die zweite Gruppe bilden die Tiff-Dateien. Hierbei handelt es sich um Bildformate, die verlustfreie Bilder enthalten können oder auch als Container für Jpeg-Bilder dienen können. Beide Varianten sind von uns geliefert worden. Die reinen Tiff-Bilder dienen als Originale, in höchster Qualität, die (weitgehend) verlustlos weiterbearbeitet und konvertiert werden können. Ebenso liegt die Tiff-Jpeg-Version bei. Für Programme als Tiff-Dateien anzusprechen sind sie wesentlich kleiner als echte Tiff-Bilder. Von Goobi wird das Tiff-Format bevorzugt, das allerdings Jpeg-Bilder enthalten darf.

Schlußendlich als dritte Gruppe die Jpeg-Bilder. Diese sind Endformate, eine Weiterbearbeitung von Jpeg-Bildern ist grundsätzlich und immer mit einem erheblichen Qualitätsverlust verbunden. Dafür sind die Dateien recht klein im Vergleich zu Tiff-Bildern.

Der DFG-Viewer verlangt Jpeg-Dateien. Er bietet vier Größendarstellungen: Thumbnails (die noch nicht realisiert sind im Viewer), klein, standard und originalgroß. Entsprechend liegen vier verschieden große Jpeg-Bilder für jede Buchseite vor.

Dazugehörig ist die vierte Gruppe: die .xml-Dateien im internen Mets-Mods-Format.

### *Sinn und Zweck der Vielfalt*

Wir liefern diese Fülle und Varietät an Dateiformaten um Ihnen alle Möglichkeiten der weiteren Nutzung und Bearbeitung der Dateien inklusive aller Präsentationsvarianten offenzuhalten und Ihnen gleichzeitig alle Qualitätsstufen anzubieten, die sinnvoll verwert- und weiterverwendbar sind.

Bei allen von uns eingesetzten Dateiformaten (Pdf, Jpeg, Tiff, Xml) handelt es sich um typische Standardformate. Das gilt gleichfalls für das verwendete Mets-Mods-Format der Xml-Dateien, das sich als derzeitiger Quasi-Standard zur Beschreibung (in Struktur- und Metadaten) von Digitalisaten etabliert hat.