Andy Stauder, Günter Mühlberger

AV-Digitalisierung zwischen zwei Stühlen

Ein Werkstattbericht zur digitalen Archivierung im Hochschulbereich

texten ein großes Problem, sondern auch in öffentlichen Einrichtungen wie Hochschulen, Bibliotheken und Archiven. Vor diesem Hintergrund und da es noch kaum finanziell tragbare Lösungsansätze für diese Problematik in diesem speziellen Szenario gibt, wurde im Rahmen des EU-Projekts *PrestoPRIME* ein Pilotprojekt an der Universität Innsbruck (siehe 2009) durchgeführt. Es geht bei dem Pilotprojekt um die Massen-Digitalisierung von AV-Medien aus dem sogenannten Consumer-Bereich, der ganz andere Charakteristiken aufweist als etwa die bei Rundfunkanstalten übliche „professionelle" Umgebung. Besonderer Wert wurde bei dem Pilotprojekt auf das Kriterium der Massendigitalisierung gelegt, da bestimmte Probleme erst bei einer großen Menge analoger Datenträger auftreten.

Einleitung und Allgemeines

Das Problem, auf welches der Titel dieses Aufsatzes anspielt, lässt sich folgendermaßen kurz skizzieren: Öffentliche Einrichtungen kleiner und mittlerer Größe – dazu zählen viele Bibliotheken und Archive, aber auch Einrichtungen wie etwa Universitätsinstitute mit ihren AV-Sammlungen – verfügen oft über substantielle und nicht selten einzigartige Medienbestände. Diese lagern zu einem großen Teil noch auf analogen Datenträgern, deren begrenzte Haltbarkeit einen der maßgeblichen Gründe für eine Digitalisierung darstellt. Eine solche Digitalisierung sollte freilich in der bestmöglichen Qualität, also möglichst getreu und ohne Verfälschung der Originale erfolgen. An dieser Stelle kommt allerdings gleich der zweite der beiden „Stühle" aus dem Titel ins Spiel: Eine möglichst hochqualitative Digitalisierung erfordert eine Menge an technischem Know-how, an Digitalisierungsgeräten sowie die Infrastruktur für die spätere Langzeitarchivierung der digitalen Daten. Ein weiteres Erschwernis und einer der wichtigsten Beweggründe für das Pilotprojekt besteht darin, dass große Teile der genannten Medien auf konsumentenorientierten Formaten lagern, welche zum einen noch stärker dem Zahn der Zeit unterliegen als professionelle Formate und welche zum anderen für die professionelle Bearbeitung nur schlecht geeignet sind, da Geräte, Workflows und Programme oftmals fehlen. Aus diesem Grund gibt es auch bislang kaum Lösungen, die auf Szenarien mit größeren Sammlungen solcher Medien zugeschnitten sind.

Um die eingangs genannten Probleme in Angriff zu nehmen, wurde das besagte Pilotprojekt der Universität Innsbruck ins Leben gerufen. Der Aufbau desselben umfasst vier Phasen. Die erste von diesen war eine Bestandsaufnahme, die über bereits bestehende Lösungen Aufschluss geben sollte. Es sollte untersucht werden, ob es im Hochschulbereich bereits erprobte Verfah-

ren gibt, um bei den Einrichtungen, die sie benutzen, Anregungen für brauchbare und erschwingliche Hard- und Software bzw. für fertige Lösungen einzuholen.

Nach den ernüchternden Ergebnissen der ersten Phase umfasste die zweite die Konzeptualisierung, den Entwurf sowie das Assembling eigener Arbeitsgeräte.

An diese schloss direkt die dritte Phase an, welche in der Auswahl, Anpassung bzw. Entwicklung der für den Arbeitsprozess benötigten Medienformate und Software bestand.

Zu guter Letzt galt es noch, die Frage nach der Datenhaltung sowie nach deren Zugänglichmachung in Angriff zu nehmen, welche auch rechtliche Überlegungen erforderte.

Pilotprojekt: Digitalisierung einer AV-Sammlung der Universität Innsbruck – Phase 1: Bestandsaufnahme

Die erste Projektphase, welche eine Bestandsaufnahme umfasste, wurde in Form einer Erhebung des Gesamtbestandes an AV-Medien der Universität Innsbruck sowie einer kursorischen Onlineuntersuchung gestaltet. Anhand letzterer Untersuchung sollte ermittelt werden, wie andere Universitäten mit den angesprochenen Problemen umgehen bzw. wie ihr Angebot im Bereich Digitalisierung aussieht. Dabei war eine erschöpfende Untersuchung allerdings nicht das Hauptziel des Projekts, da davon auszugehen war, dass eine solche einen beträchtlich größeren, v.a. zeitlichen Aufwand bedeuten würde, als es zu dem Zeitpunkt zweckmäßig gewesen wäre. Eine solche umfassende Untersuchung, die in einer aktiven Onlineumfrage besteht, die an annähernd 800 europäische Hochschulen gerichtet wird, ist allerdings derzeit in Vorbereitung und soll die allgemeine Situation von AV-Medien an europäischen Hochschulen erheben.

Was die Erhebung des Gesamtbestandes an der Universität Innsbruck anbelangt, so konnten im Jahr 2009 an den 20 Instituten mit nennenswerten Beständen insgesamt etwa 60.000 Datenträger mit ungefähr 82.000 Stunden an Material verzeichnet werden: also eine beträchtliche Menge an (teilweise stark bedrohtem) Material.

Hinter der Online-Untersuchung stand die Überlegung, dass Universitäten, so sie über ein effizientes Digitalisierungsangebot bzw. digitalisierte Bestände verfügen sollten, dieses wohl im Rahmen ihrer Internetpräsenzen darstellen oder bekannt machen würden. Dementsprechend wurden mehrere ausgewählte Begriffe für eine Suchmaschinenabfrage verwendet und die relevantesten Treffer unter den Universitäts-Websites einer genaueren Betrachtung unterzogen. Diese Universitäten bestanden in den folgenden: Universität Regensburg; Cape Breton University, Canada; Indiana University, USA; Friedrich-Schiller-Universität Jena; Universität Rostock; Philipps-Universität

Marburg; Universität Würzburg; Universität Graz; Universität Wien; University of Oxford, Großbritannien.

Die Ergebnisse zeigten dabei eine recht einheitliche Tendenz: Meist gab es nur Digitalisierungsangebote für Print-Material. Die AV-Digitalisierung bzw. Verlagerung auf migrierbare Medien schien weitgehend vernachlässigt. Dies ist v.a. aus einem Grund befremdlich: Datenträgergenerationen sind in der Regel umso empfindlicher, je jünger sie sind (vgl. u.a. Friedewald/Leimbach 2011: 204ff.), während die Datenmenge pro Fläche auf diesen gleichzeitig zunimmt, wobei Ersteres der Hauptverursacher von Letzterem ist.

Ein ebenfalls recht einheitliches Bild zeigt sich, wenn man den operativen Aspekt der Digitalisierungsangebote betrachtet: An den untersuchten Universitäten sind die Möglichkeiten zum Digitalisieren von audiovisuellem Material in der Regel „nachfragebasiert". Das bedeutet, sie sind entweder „zum Selbermachen", mit einer Anleitung oder kurzen Einschulung an den vorhandenen Geräten, oder sie bestehen in einem Bring-Hol-Schema, wobei man die Datenträger abliefert und die Digitalisate dann wieder abholen kann. In beiden Fällen ist dabei die Menge der verarbeiteten Datenträger pro Zeit gering, und daher sind diese Arten der Verarbeitung eher für einzelne Stücke oder kleinere Stückzahlen geeignet.

Dieser Umstand ist v.a. dem Aufbau der Maschinen geschuldet. Das Schema, das dabei meistens zur Anwendung kommt, ist A → B. Das heißt, genau ein Abspielgerät für einen analogen Datenträger ist an genau einen Computer mit A/D-Wandlerhardware angeschlossen. Varianten sind dabei, dass das Abspielgerät zugleich die Digitalisierung vornimmt, wie etwa bei Festplattenrecordern, oder dass mehrere Abspielgeräte an einen Computer angeschlossen sind, aber nicht zugleich verwendet werden können. Dieses Schema sieht dann folgendermaßen aus:

(A xor B xor C … xor N) → X[1].

Die Speicherung und Zugänglichmachung des Materials war ebenfalls in keinem der untersuchten Fälle so gestaltet, wie es für das Projekt angestrebt war. Dieser Schritt erfolgte auf tragbaren Datenträgern, die den Benutzern selbst gehörten, oder auf gebrannten optischen Datenträgern. Eine zentrale Lösung wie ein Repositorium sowie ein Zugriff auf die Daten via Streaming bzw. die rechtliche Regelung einer solchen Möglichkeit fehlte in allen untersuchten Fällen.

Aufgrund dieser Ergebnisse wurde die Entscheidung getroffen, eine hausinterne Sammlung für das zu untersuchende Szenario (also preisgünstige Massendigitalisierung von audiovisuellen Inhalten im öffentlichen bzw. vor allem im Hochschulbereich) heranzuziehen. Die Wahl fiel dabei aufgrund mehrerer Eigenschaften, welche für das Untersuchungsszenario typische Probleme dar-

[1] <xor> bezeichnet ein ausschließendes „Oder".

stellen, auf die Videosammlung des Instituts für Slawistik. Eine dieser Eigenschaften ist die Sammlungsgröße. Diese macht mit 3.000 physischen Objekten zahlenmäßig etwa fünf Prozent des Gesamtbestandes der Universität aus, und da viele der Datenträger eine Spielzeit von vier Stunden haben, stellt die Sammlung bezüglich der Menge an Inhalt mit über 7.000 Stunden annähernd zehn Prozent des Gesamtbestandes. Was die Formate anbelangt, so enthält die Sammlung ca. 1.800 VHS-Kassetten, 1.100 DVDs und 74 Rollen Schmalspurfilm. Die Sammlungsgröße ist v.a. aus dem Grund interessant, dass sie für das untersuchte Szenario typische Objektzahlen mit sich bringt. Diese haben es zur Folge, dass eine solche Sammlung nicht mehr mit Schemata wie A → B oder (A xor B xor C … xor N) → X in einem Zeitraum von unter einem Jahr bearbeitet werden können.

Dies ist dadurch bedingt, dass analoge Datenträger nicht wesentlich schneller als mit einfacher Geschwindigkeit abgespielt werden können, ohne dass sich merkliche Qualitätsverluste einstellen: Ein beschleunigtes Abspielen verändert die physikalischen Rahmenbedingungen der Abtastung, was eine Veränderung des Ausgangssignals zur Folge hat. Bei digitalen Medien liegen die Inhalte abstrahiert in Zahlenform vor, wodurch Kontrollmechanismen (wie Prüfsummenverfahren) das Erfassen aller Inhalte beim Abtasten mit großer Sicherheit überprüfen können. Genauer gesagt: Es wird das physikalische Signal auf dem Datenträger diskretisiert (also in Form von genauen Zahlenwerten interpretiert), wobei auch Kontrollelemente auf dem Medium vorhanden sind, die, nachdem sie ebenfalls diskretisiert worden sind, darüber Aufschluss geben, ob jene Teile des Signals, für welche sie verantwortlich sind, richtig als Zahlen interpretiert worden sind. Dadurch kann immer bestimmt werden, ob noch ein Nutzsignal ausgelesen wird oder nicht, und auf diese Weise sinnvolle Abspielgeschwindigkeiten bestimmt werden. Im Fall der Analogsignale würden nur immer feinere Detektoren (z.B. Magnetköpfe) helfen, welche schnell sehr teuer werden würden. Außerdem bedeutet eine beschleunigte Abspielgeschwindigkeit eine erhöhte Beanspruchung des Datenträgers, was bei zu sichernden Analogmedien ein zusätzliches und nicht zu vertretendes Risiko darstellt.

Es ist also aus Gründen der Praktikabilität ein Abspielen mit Normalgeschwindigkeit für die Digitalisierung erforderlich. Das bedeutet wiederum, dass mit den o.g. Schemata durch eine vollzeitbeschäftigte Person nur eine sehr begrenzte Menge an Material digitalisiert werden kann. Außerdem sind die Speicherplatzmengen für derartige und größere Mengen an Material ebenfalls nicht zu vernachlässigen, umso mehr, wenn man eine Zurverfügungstellung, z.B. via Streaming, ins Auge fasst, welche den besonders teuren hochverfügbaren Speicher benötigt.

Eine weitere interessante Eigenschaft der Sammlung ist, dass alle Objekte über sehr detaillierte deskriptive Metadatensätze in einer Web-Datenbank verfügen. Diese wollen verwaltet und für die digitale Nutzung der jeweiligen

Medien angepasst werden. So galt es unter anderem, die Metadaten, die in einem vom Institut für Slawistik zusammengestellten Format vorliegen, zu standardisieren, damit die digitalisierten Objekte später im Rahmen einer digitalen Bibliothek an der Universität genutzt werden können. Es ging also um ein Metadaten-Mapping. Außerdem bringen die in diesem Fall vorliegenden Metadaten noch eine weitere Anforderung mit sich: Der Inhalt der Datensätze liegt zum Großteil in slawischen Sprachen vor, weswegen die entsprechenden Schriftsätze berücksichtigt werden müssen.

Ebenfalls eine interessante Eigenschaft ist das Vorhandensein verschiedener Datenträgerarten: VHS mit unterschiedlichen Farbformaten (PAL und SECAM), DVD (eigentlich kein analoges Medium, aber eines, das trotzdem gesichert werden muss; mit der Schwierigkeit, dass dies ein hierarchisch strukturiertes Medium ist) und Schmalspurfilm (mit der Schwierigkeit, dass qualitativ hochwertige Digitalisierungsgeräte hierfür schwer verfügbar sind).

Phase 2: Assembling der Hardware

Die zweite Phase umfasste die Konzeptualisierung einer Hardwarelösung, welche die in Phase 1 identifizierten Probleme zu lösen helfen sollte. Dabei sollte ein Konzept entstehen, das flexibel, günstig und einfach zu handhaben ist. Zu diesem Zweck wurde ein System mit einem leistungsstarken Server entwickelt, an welchen zugleich verschiedene oder auch gleichartige Abspielgeräte angeschlossen werden können. Auf diese Weise konnten mehrere Vorteile erzielt werden: Es konnten Kosten eingespart werden, die bei einem Modell mit mehreren Rechnern unnötig entstehen würden (weil viele Komponenten dabei mehrfach vorhanden wären, aber nicht voll ausgenutzt würden, z.B. Netzteile, Gehäuse usf.). Außerdem ließ sich so ein geringerer Platzbedarf bzw. Mobilität erreichen. Ebenfalls konnte die Steuerung sehr einfach gehalten werden.

Da das System so ausgelegt ist, dass beliebige Abspielgeräte angeschlossen werden können, konnte man sich bei diesem ersten Pilotprojekt auf eine einzige Art von Geräten beschränken, da eine Erweiterung später ohnehin möglich wäre. Es wurde dabei das am meisten bedrohte Medium gewählt: die VHS-Kassetten, welche außerdem noch den Großteil der Sammlung ausmachen.

Doch nun zum System selbst: Die grundlegende Komponente des zentralen Rechners ist das Mainboard. Hier wurde eine Server-Variante mit den benötigten Eigenschaften gewählt: zwei Sockel für Vierkern-CPUs, welche die Grundlage für ausreichend Leistung bieten sollten, um 6-8 Streams zugleich zu bewältigen; die Leistung wurde dabei von Desktoprechnern extrapoliert, mit denen Vorlauftests durchgeführt wurden; außerdem sollte die Hauptplatine genügend Steckplätze für Erweiterungen – die A/D-Wandler – bieten.

Weitere wichtige Komponenten waren dann die CPUs, wobei hier Vierkernmodelle mit 2,13 GHz gewählt wurden, sowie die vier Hauptspeichermodule

(DDR3 mit 1333 MHz) à 2 GB und der Plattenspeicher von acht einzelnen, je ein TB fassenden Laufwerken, welche das bei AV-Bearbeitung zu erwartende hohe Datenaufkommen bewältigen sollten.

Die Komponenten, die den Hauptzweck des Gerätes erfüllen sollten, sind natürlich die A/D-Wandler. Hier wurden zunächst USB-basierte Modelle verwendet, welche allerdings trotz ihrer akzeptablen Leistung nicht alle getesteten Qualitätsanforderungen erfüllen konnten. Daher fiel zuletzt die Wahl auf PCI-express-basierte Wandler aus dem professionellen Studiobereich. Von diesen konnten zum Zeitpunkt dieser Publikation sechs hochqualitative Digitalisierungsvorgänge gleichzeitig durchgeführt werden, was also bedeutet, dass im Vergleich zu einem oben genannten Schema die Arbeit eines Jahres in 2 Monaten verrichtet werden kann, wobei die Hardwarekosten nur ungefähr das 3-4-Fache betragen, d.h. gerechnet auf die Verarbeitungskapazität pro Zeit ca. die Hälfte bis zwei Drittel des Preises anfallen; wenn man mit einem Modell mit mehreren Einzelrechnern vergleicht, welches denselben Durchsatz hätte, wird die Kostenersparnis durch das Einmaschinen-Modell angesichts des Mehrs an benötigten Monitoren, Betriebssystemlizenzen und Eingabegeräten des Mehrmaschinen-Modells ebenfalls deutlich: Letzteres würde etwa das Doppelte kosten. Außerdem ist die entwickelte Lösung zusätzlich fahrbar, denn das ganze System ist auf einem Rollwagen montiert und benötigt so auch nur bis zu einem Sechstel des Platzbedarfs.

Phase 3: Formate/Software

In der dritten Phase war in erster Linie die Wahl eines brauchbaren Formats für die Daten von Bedeutung. Hier fiel die Entscheidung letztendlich auf MP4 als Containerformat mit mp4v als Videocodec und aac als Audiocodec. Hier fällt als erstes auf, dass dies ein verlustbehaftet komprimiertes Format ist. Diese Entscheidung wurde aus der Überlegung heraus getroffen, dass ein unkomprimiertes oder auch verlustfrei komprimiertes Format von den anfallenden Speicherkosten im untersuchten Szenario schlichtweg nicht tragbar ist: Diese würden nämlich das 30-Fache betragen, was auch bei dem relativ günstigen Magnetbandspeicher in den zeitgenössischen Rechenzentren noch schwer ins Gewicht fällt. Allerdings ist die Maschine so ausgelegt, dass auch eine unkomprimierte oder verlustfrei komprimierte Produktion der Daten möglich ist.

Jedenfalls wurden für die verwendete Komprimierung die Qualitätseinstellungen mit einer Auflösung von 720x576 Pixel bei einer Bandbreite von 5,5 MBit/s[2] so gewählt, dass die Inhalte des ohnehin informationsarmen Mediums VHS (manche Experten halten z.B. die verwendete Auflösung bereits

[2] Das bedeutet ein Datenaufkommen von etwa 2,5 GB pro Stunde Spielzeit. Im Vgl. dazu würde eine unkomprimierte Datei 72 GB/Std. benötigen, wenn man sie verlustfrei komprimiert, etwa 30-36 GB/Std.

für mehr als ausreichend: siehe Johnson und Crawford 2006) unter normalen Gebrauchsbedingungen ohne merklichen Qualitätsverlust wiedergegeben werden kann: Das heißt, die Anzeige eines Digitalisats auf einem 20-Zoll-Monitor aus einem Meter Betrachtungsabstand ist durch einen Menschen qualitativ nicht von einer Anzeige des Originalinhalts auf demselben Monitor zu unterscheiden.

Was die Software anbelangt, so wird mit einer Kombination aus kommerziellen, Open-Source- und selbstentwickelten Lösungen gearbeitet. Die Datenträger (in diesem Fall die VHS-Kassetten) werden für die Digitalisierung manuell in die angeschlossenen Videorekorder eingelegt, zurückgespult und dann mit der Softwarekombination digital aufgezeichnet. Pro Charge bedeutet das einen Aufwand von maximal 10 Minuten. Hierbei ist die Entwicklung des Workflows aber trotz vollständiger Funktionstüchtigkeit noch nicht vollständig abgeschlossen, weswegen an dieser Stelle noch kein detaillierter Bericht darüber abgegeben werden kann.

Phase 4: Datenhaltung/Zugänglichmachung

Der vierte Teil des Pilotprojekts befindet sich noch in der Planungs- und Entwicklungsphase. Hierbei wird angestrebt, ein zentrales Repositorium für die Daten aus der Digitalisierung sowie die jeweils zugehörigen Metadaten zu schaffen. Dabei soll zumindest für den Zugriff auf die Daten durch die Benutzer die Plattform eingesetzt werden, die aus den gemeinschaftlichen Anstrengungen der Projektpartner von PrestoPRIME hervorgegangen ist. Diese „PrestoPRIME Preservation Platfom" (für eine genauere Beschreibung siehe Gallo 2011), kurz P4, ist als Open-Source-Entwicklung realisiert und umfasst eine Reihe von Modulen, die für den Betrieb eines AV-Repositoriums benötigt werden. Die Universität Innsbruck hat dabei einen auf der Lucene-Software basierenden Suchdienst für den benutzerfreundlichen Umgang mit den Metadaten beigesteuert. Was die Metadaten selbst anbelangt, so hat jedes Videofile entsprechend der Kassette, zu der es gehört, eine ID-Nummer als Dateinamen. Über diese Nummer können den verschiedenen Inhalten der Datei – viele Kassetten waren mit mehreren inhaltlichen Objekten bespielt – die passenden Metadatensätze aus einem Auszug der Metadatenbank zugeordnet werden. Auf diese Weise ist ein manuell aufwändiges bzw. maschinell unsicheres (d.h. viel Überwachung benötigendes) Schneiden in einzelne Dateien zunächst nicht notwendig: Alle Objekte können auch so gefunden und verwendet werden. Das Schneiden wird später ohne Zeitdruck nach und nach manuell durchgeführt und in der digitalen Sammlung schrittweise die Objektbündel mit Einzelobjekten ersetzt.

Die Daten selbst sollen zum einen für die Langzeitsicherung als sogenannte Masterfiles auf Bandspeicher abgelegt werden und zum anderen als Gebrauchskopien im Webm-Format via HTML-5-Technologie an speziellen Arbeitsplätzen auf dem Universitätsgelände verfügbar gemacht werden, wobei

die Zahl der möglichen simultanen Zugriffe auf jeweils ein Objekt begrenzt werden muss, um die urheberrechtliche Unbedenklichkeit zu gewährleisten.

Fazit

Die Erhaltung audiovisueller Ressourcen aus dem analogen Zeitalter ist zum einen mit relativ großem Aufwand verbunden und erfordert in Fällen, wo größere Mengen betroffen sind, einen beträchtlichen Mehraufwand, welcher v.a. in mittelgroßen und/oder strukturell nicht selbstständigen Einrichtungen wie Bibliotheken oder Universitätsinstituten finanziell, logistisch und bezüglich der Qualität selten geleistet werden kann. Das vorgestellte Pilotprojekt hat versucht, aus allen zur Verfügung stehenden Optionen die praktikabelsten auszusuchen und ein gangbares Modell zu schaffen, welches soviel wie möglich Anforderungen so weit als möglich bei möglichst geringem finanziellen Aufwand erfüllt. Der wichtigste Grundsatz war dabei: Nicht das teuerste Modell gewinnt, sondern dasjenige mit der besten Balance zwischen Qualität und Praktikabilität, denn die Inhalte auf den alten Datenträgern warten nicht, bis eine Lösung erschwinglich wird.

Quellen

Friedewald, Michael/Leimbach, Timo (2011). „Computersoftware als digitales Erbe. Probleme aus Sicht der Technikgeschichte". In: Robertson-Von Trotha, Caroline/Hauser, Robert (Hg.). Aspekte, Perspektiven und Konsequenzen der Digitalen Überlieferung. Karlsruhe.

Gallo, Francesco (2011). Projekt PrestoPRIME: Deliverable 5.2.2. „First Prototype of Open PrestoPRIME Reference Implementation". https://prestoprimews.ina.fr/public/deliverables/PP_WP5_D5.2.2_FirstPrototype_R0_v1.00.pdf (Stand 04.01.2012).

PrestoPrime-Projekt (2009). Projekt-Website. URL: http://www.prestoprime.org/project/index.en.html (Stand 04.01.2012.)