Emulation, Migration, Normalisierung

Langzeiterhaltung digitaler Inhalte für das CultLib-Projekt I

Wir sichern Kulturgut in digitaler Form
und machen es frei verfügbar,
heute und für zukünftige Generationen.

CultLib-ID 00aff142-045a-46ef-aa27-21ce3f8ef057
Version: 1.0
Datum: 9. April 2014
Verfasser: Dr. sc. math. Hartwig Thomas

Übersicht

Das Projekt *CultLib* des Vereins *Digitale Allmend* bezweckt die Gründung einer schweizerischen Stiftung *Pro Cultura Libera* welche ein digitales Repositorium freier, kultureller Werke aufbaut und betreibt.

Die Berichte über Aspekte der Langzeiterhaltung[1] erklären, wie CultLib technisch konzipiert ist und dienen als Anforderungen an die Umsetzung. Ausserdem können sie für Entscheidungsträger und IT-Architekten von Memo-Institutionen von Interesse sein.

In diesem ersten Bericht über Erhaltungsstrategien wird dargelegt, dass das zentrale Problem der Langzeiterhaltung nicht die Lebensdauer der Datenträger sondern die Wahl der Datenformate ist.

Emulation und Migration, die traditionellen Ansätze der Langzeitarchivierung, sind ungenügend. Sie müssen ergänzt werden durch Normalisierung. Damit rückt die Frage der für die Langzeiterhaltung geeigneten Datenformate in den Vordergrund. Proprietäre Formate eignen sich dafür grundsätzlich nicht.

Bei der Wahl der Formate stellen sich Fragen nach offenen Standards und Fragen nach der unterschiedlichen Qualität von verschiedenen Ausprägungen eines digitalen Dokuments.

[1] Der Begriff Nachhaltigkeit wird im Zusammenhang mit der Langzeiterhaltung bewusst vermieden, weil die virtuelle Allmend in ihrer Ausdehnung nicht beschränkt ist, und deshalb die mit dem Begriff Sustainability adressierte Übernutzung, die traditionelle Tragödie der Allmend, im Bereich der digitalen Inhalte keine Gefahr darstellt. Das Ziel der Tradierung kultureller Inhalte ist würdig und gewichtig genug und hat nicht nötig, sich auf die Leere übernutzter Warmluft-Begriffe zu stützen.

Inhaltsverzeichnis

1 Neue Herausforderungen bei der Archivierung digitaler Inhalte

Die wichtigsten Herausforderungen der Archivierung *analoger* Inhalte sind:

- Platzprobleme,
- Tektonik des Archivs,
- Katalog, Signaturen und Magazin,
- Erhaltungsmassnahmen verfallender Archivinhalte,
- Verluste ausgeliehener Materialien.

Auf den ersten Blick scheinen die Verhältnisse bei digitalen Archiven[2] ähnlich zu liegen:

Auch der Speicherplatz digitaler Inhalte nimmt laufend in dem Masse zu, in welchem das Archiv neue Inhalte übernimmt. Der Aufwand für die hierarchische Strukturierung, der Beschlagwortung und der Thesauren der Unterlagen scheint ins Unermessliche zu wachsen. Es ist schwierig, die wahren Standorte (Dateinamen) und diejenigen im Katalog (Datenbank), der die Metadaten enthält und zur Recherche dient, synchron zu halten. Dateiformate, Abspielsoftware und -geräte werden so schnell obsolet, dass Dateien nicht mehr maschinell gelesen werden können, womit ein Informationsverlust droht.

Auf den zweiten Blick, stellen sich die Probleme allerdings doch etwas anders. Die Platzprobleme betreffen Speicherplatz und nicht physischen Platz. Speicherplatz beansprucht immer weniger physischen Platz. Nicht der Raumbedarf, sondern die Strukturierung des Magazins ist eine neue Herausforderung. Während man früher „bewertete“, was sich ins Archiv aufzunehmen lohnt, ist es heute oft lohnender, keine teure Bewertung vorzunehmen, weil Speicherplatz billiger ist als die Bewerter. Die Tektonik des Archivs, der richtige hierarchische Thesaurus, die Beschlagwortung verlieren an Wichtigkeit, wenn man das Archiv auch mit automatischen Erschliessungsmitteln – etwa mit Volltextsuche – durchstöbern kann. Eine solche ist nicht immer ein voller Ersatz für einen klugen hierarchischen Zugriff. Wenn aber Beschlagwortung und Thesaurus wertvolle Arbeitskraft kosten, ist die automatische Erschliessung eine wichtige Alternative, die ausserdem direkt auf den authentischen Primärdaten beruht und keine Kanalisierung des Interesses der Archivnutzer zulässt, wie sie bei menschlicher Erschliessung unvermeidlich ist.

Auch die physikalische Haltbarkeit der Trägermedien ist nicht in gleicher Weise ein Problem wie bei der Archivierung analoger Inhalte. Es mag sein, dass selbstgebrannte CD nach einigen Jahrzehnten nicht mehr lesbar sind. Auch Disketten und Magnetbänder haben oft nach Jahrzehnten physisch ihre Magnetisierung verloren. Ausserdem sterben die Lesegeräte für solche Datenträger aus. Aber ein Archiv wird kaum seine Unterlagen auf physikalischen Trägern in einem physikalischen Magazin im Keller lagern und diese – wie früher die analogen Datenträger – nur dann hervorholen, wenn ein Archivnutzer nach ihnen verlangt. Stattdessen wird es den gesamten Bestand im Direktzugriff auf Datenträgern der aktuellen Speichertechnologie in mehreren Kopien speichern und diesen Bestand auf die jeweils neue aktuelle Speicher-

[2] Wo im Folgenden von Archiven die Rede ist, sind ganz allgemein Memo-Institutionen wie Bibliotheken, Museen und eigentliche Archive gemeint; das Wort Archivierung wird synonym für Erhaltung oder Aufbewahrung benutzt.

technologie migrieren, denn beim Kopieren digitaler Inhalte geht kein einziges Bit verloren. Die Dateien und ihre Verzeichnisse sind nach der Migration gleich strukturiert wie vorher.

Schliesslich ist auch das Risiko des Verlusts „ausgeliehener“ Unterlagen verschwunden. Denn Unterlagen werden nicht „ausgeliehen“, sondern kopiert. Während ein Archivnutzer eine angeforderte Unterlage liest, bleibt sie im Magazin für Andere verfügbar. Hier stellt sich höchstens das umgekehrte Problem einer Verfügbarkeit, die eventuell über das erwünschte Mass hinausgeht, wenn Schutzfristen und andere Auflagen den Nutzern eine weitere Verbreitung verbieten.

2 Eine Frage der Formate

Die Langzeitarchivierung digitaler Inhalte scheint also zentral mit der Frage der Formate verbunden zu sein. Es ist einfach, einen Bit-Strom mittels Kopieren auf einen anderen Träger zu migrieren. Es ist aber schwierig bis unmöglich, ein archiviertes Bild zu lesen, wenn es in einem undokumentierten Format archiviert wurde, wie etwa dem Photo-CD-Format, das von KODAK geheimgehalten wurde, und heute nach dem Konkurs dieser Firma von keiner Software mehr korrekt interpretiert werden kann.

2.1 Emulation, Migration und Normalisierung

Der drohenden Obsoleszenz von Datenformaten und Abspielgeräten wird traditionellerweise mit den Methoden der *Emulation* oder der *Migration* begegnet.

2.1.1 Emulation

Das Konzept der Emulation ist schon auf mechanisch abspielbare analoge Datenträger anwendbar. Man benutzt zum Abspielen einer alten Schellackplatte ein Gerät, welches den originalen Plattenspieler imitiert. Dieses analysiert zum Beispiel die dreidimensionale Struktur der Rillen in der Platte mit Laserstrahlen und setzt sie in einen Ton um, wie ihn das alte Grammophon erzeugt hätte.

Auch in der digitalen Welt können ältere Maschinen auf neueren emuliert, d.h. in ihrer Funktionsweise nachgeahmt werden. So gibt es etwa unter modernsten Betriebssystemen Emulationen von DOS, Amiga und Atari. Das Konzept der Emulation für die Langzeiterhaltung beruht aus dieser Idee: Statt Dokumente in obsoleten Formaten (z.B.: eine WordPerfect-Datei) als unlesbar zu erklären, wird einfach ein Exemplar von WordPerfect in einer Emulation des Betriebssystems gestartet. Mit dieser Abspielsoftware kann dann das Dokument gelesen werden.

Dieser Ansatz ist mit zwei grossen Problemen verknüpft. Damit ein System (z. B. das Betriebssystem oder WordPerfect) emuliert werden kann, muss seine Spezifikation frei und offen vorliegen. Ein KODAK-Photo-CD-Emulator kann nicht hergestellt werden, wenn sein Aufbau nicht frei und offen zugänglich ist oder mindestens das Betriebssystem, auf dem die KODAK-Software zum Einsatz kommt frei und quelloffen zur Verfügung steht. Dies dürfte in den wenigsten Fällen gegeben sein. Zweitens ist es unwahrscheinlich, dass spätere Archivnutzer fähig sind, ein vorsintflutliches Word oder Wordperfect zu bedienen. Auch die Bedienungsanleitungen der Geräte und Systeme müssten archiviert werden. Wenn man in Betracht zieht, wie hoch das Salär eines Datenbankadministrators heute ist, kann man sich vorstellen, was es kosten wird, wenn man einen Datenbankadministrator für die Bedienung der Emulation von vor fünfzig Jahren aktuellen Datenbanken benötigt.

2.1.2 Migration

Unter Migration versteht man das laufende „Umkopieren“ nicht nur der Bits sondern auch der Formate. Man „migriert“ also eine DOS-Word-Datei ins Win-

Word-2-Format und dann auf WinWord 6 und dann auf Office 2003 und dann ins DOCX-Format auf Office 2007 und dann auf Office 2010 und so weiter. Das Problem bei solchen Migrationen liegt darin, dass die „Authentizität“ verloren geht. Nach der Migration sieht das Archivgut völlig anders aus. Die Qualität der archivierten Unterlagen ist teilweise eine Funktion der Formate. Es geht immer etwas verloren und es wird immer etwas hinzugefügt. Wenn man nicht die Originale *und* den Quellcode der Migrationssoftware aufbewahrt, ist der Bezug zum Original nicht mehr nachvollziehbar.

Dumme Programmfehler in der Migrationssoftware, wie sie etwa bei der Migration von Dokumenten mit Diagrammen von Office 2003 zu Office 2007 vorhanden sind, könnten beträchtliche Teile des Archivs zerstört haben, ohne dass man das rechtzeitig merkt. Im obigen Beispiel mit Microsoft Word dürfte eine Datei aus DOS-Zeiten kaum wohlbehalten im Office 2010 angekommen sein. Da die Migrationssoftware von Microsoft nicht quelloffen ist, kann man nicht einmal nachvollziehen, wo wann was verloren ging. Die Strategie der Migration ist aber gerade der Tatsache geschuldet, dass das Datenformat, die Abspielsoftware und die Migrationssoftware proprietär ist. Wenn das Datenformat frei und offen standardisiert vorläge, wäre keine Migration nötig.

2.1.3 Normalisierung

Das Konzept der Normalisierung ist im Archivbereich leider noch nicht sehr verbreitet. Es basiert gerade darauf, dass bei freien und offenen Formaten eigentlich keine Migration notwendig ist. Das beobachtet man etwa beim einfachsten und ältesten Format der einfachen Textdatei, welche eine rohe Zeichenfolge speichert. Während eine Worddatei aus DOS-Zeiten heute nicht mehr brauchbar ist, merkt man einer einfachen Textdatei praktisch nicht an, ob sie zehn, dreissig oder fünfzig Jahre alt ist. Es gibt einige wenige offene Standards, wie die Buchstaben in Bytes zu kodieren sind (ASCII, EBCDIC, ISO Latin, UTF-8, ...). Im Kernbereich der unverzierten Buchstaben des Alphabets sind sie leicht ineinander übersetzbar. Einzig, dass sie keine Identifikation ihres Formats enthalten , kann gewisse Schwierigkeiten bereiten.

Wenn wir ein Bild in irgendeinem Format abspeichern, das frei und offen und klar so standardisiert ist, dass man zu jeder Zeit aus dem Bit-Strom eruieren kann, welche Farbe an welchem Ort abzubilden ist, kann man zu jeder Zeit eine neue „Abspielsoftware“ herstellen, welche das Bild korrekt wiedergibt, solange wir Zugriff auf den Standard haben. Dabei ist es nicht so wichtig, ob das Archivformat zum Zeitpunkt der Nutzung noch nennenswerte Verbreitung hat. Man kann es zu jedem Zeitpunkt in ein dannzumal verbreitetes Format konvertieren, es denormalisieren. Selbst zum Zeitpunkt der Archivierung braucht das Format kein verbreitetes Format zu sein. Es muss ja keine Abspielgeräte speziell gut unterstützen, sondern sich nur für die Langzeitarchivierung eignen. Bei der Übernahme in das Archiv kann ein Bild aus einem aktuell gerade verbreiteten Format in das Normalformat konvertiert, normalisiert werden.

Die Archivierungsstrategie der Normalisierung lässt sich also so zusammenfassen:

- Bei der Übernahme eines Dokuments wird dieses aus dem *zu dieser Zeit* populären Format in ein Normformat umgewandelt. Es wird normalisiert.
- Dieses normalisierte Dokument bleibt für immer unverändert im Archiv.
- Bei der Vermittlung des Dokuments wird es aus dem Normformat in das *zu jener Zeit* populäre Format umgewandelt. Es wird denormalisiert.

Der entscheidende Punkt ist, dass das normalisierte Dokument im Archiv nie mehr verändert zu werden braucht und somit seine Authentizität besser gewährleistet werden kann.

Die Prozesse der Normalisierung und der Denormalisierung sind von der aktuellen Technologie abhängig. Nicht das Archivgut wird migriert, sondern die Prozesse der Normalisierung und Denormalisierung. Das normalisierte Archivgut kann hingegen beliebig lange unverändert aufbewahrt werden. Mit „unverändert“ meinen wir nicht, dass es während der ganzen Zeit auf demselben materiellen Träger gespeichert ist, sondern dass die Datei, der Bit-Strom, der die Unterlage repräsentiert, unverändert bleibt.

Ob sich ein Format als Normalformat eignet, erkennt man an folgenden Eigenschaften:

1. Die Formatbeschreibung ist ein freier, offener, internationaler Standard.
2. Das Format ist nicht auf spezielle Abspielgeräte ausgelegt oder von ihnen abhängig.
3. Das Format deckt den ganzen Bereich der darzustellenden Objekte ab.
4. Das Format kann für Objekte verschiedener Qualität verwendet werden.
5. Im Format abgespeicherte Objekte enthalten einen Hinweis auf das verwendete Format.
6. Das Format unterstützt eingebettete Metadaten.

2.2 Freie, offene, internationale Standards

Es besteht ein Konsens, dass freie und offene Standards eine wichtige Grundlage für die Weiterentwicklung des Geisteslebens sind. Es gibt sehr viele Vorschläge für eine Definition dessen, was als offener Standard zu gelten hat. Siehe etwa http://en.wikipedia.org/wiki/Open_standard. Am nützlichsten scheinen mir kurze, pragmatische Definitionen wie etwa diejenige der Free Software Foundation: http://fsfe.org/activities/os/def.html.

Aus der Archivsicht ist es nicht so wichtig, wie der Standardisierungsprozess verlaufen ist. Es ist auch nur von zweitrangiger Bedeutung, ob ein Standard von einer privaten Firma entwickelt wurde oder von einem internationalen Gremium abgesegnet wurde. (Für konkurrierende Firmen, die sich auf eine gewisse Interoperabilität ihrer Systeme einigen sollen, wollen oder müssen, sind diese Fragen natürlich von eminenter Bedeutung.) Der ZIP-Standard

wurde zum Beispiel von einer privaten Firma (Philip Katz, PKWARE Inc.) entwickelt und offengelegt und nie von einer internationalen Organisation heiliggesprochen, obwohl er unter anderem dem „freien und offenen“ ISO-Standard Open Document Format zugrundeliegt. Allerdings hat PKWARE den Standard frei und offen publiziert und in die Public Domain übergeben. Bis auf einige, inzwischen weitgehend abgelaufene Patente auf gewissen Kompressions- und Verschlüsselungsalgorithmen ist der ZIP-Standard weltweit frei benutzbar für Alle. Es existieren viele miteinander interoperable Implementationen, es existiert ein Standard-Dokument, auf dessen Basis man neue Implementationen programmieren kann. Die Benutzung des Standards hängt von keiner proprietären Hardware oder unfreien Software ab. Der Standard ist weit verbreitet und wird in den nächsten fünfzig Jahren wohl kaum vergessen gehen. (Garantierte Verfügbarkeit und Verständlichkeit von Archivgut nach fünfzig Jahren ist eine Mindestanforderung an „Langzeit“-Archivierung.)

Die wichtigste Funktion der internationalen Standardisierung eines Dateiformats besteht aus der Archivperspektive darin, dass diese sicherstellt, dass der Standard nicht mit der Zeit verloren geht. (Bei Schriften, welche ja auch schon einen „digitalen Standard“ darstellen, hat die Menschheit bisher mehrheitlich den Verlust ihrer Bedeutung bravourös zu vermeiden gewusst.) Ein offener Standard muss in vielen Kopien, möglichst auf verschiedenen Medien (inkl. auf Papier in Bibliotheken) und in verschiedenen Formaten weltweit frei verfügbar sein. Durch solche Redundanz wird die Langzeiterhaltung aller Dokumente im standardisierten Datenformat gesichert.

Manchmal reicht auch das schiere Volumen der in einem speziellen Format vorhandenen Daten aus, um einen Standard festzuschreiben. Das CD-Audio-Format („Red Book“-Format) wurde ursprünglich von einem Industriekonsortium (Philips, Sony) festgelegt, welches anfänglich den Zugang zur Spezifikation absichtlich erschwerte. Auch wenn später internationale Standards (IEC 908 / IEC 60908) festlegten, wie eine CD-Audio strukturiert zu sein habe, sind die Standardisierungsdokumente nur schwer zugänglich. Wer eine Kopie des „Red Book“ haben möchte, muss 100$ an Philips zahlen und eine Vertraulichkeitserklärung unterschreiben. Wenn Philips – wie KODAK – einmal Konkurs gehen sollte, wird aber der CD-Audio-Standard nicht verloren gehen, denn es gibt unzählige Beschreibungen des Formats, weil es so viel Musik in diesem Format gibt.

2.3 Vermeidung von Vendor-Locking

Von den weltweit hundert grössten Firmen im Jahr 1900 existierten hundert Jahre später noch zwei. Langzeiterhaltung kann sich also nicht auf vermutete Firmenerhaltung abstützen, insbesondere die Mehrheit der Firmen über den Konkurs hinaus keinerlei Haftung unterworfen ist. Niemand hätte 1995 geglaubt, dass es die Swissair in zehn Jahren nicht mehr geben wird. Niemand hätte es 2005 für möglich gehalten, dass eine Weltfirma mit Tradition wie KODAK zehn Jahre später nicht mehr existiert. Niemand glaubt heute, dass es Oracle, Adobe, Apple, Microsoft, Google, etc. in zehn Jahren nicht mehr geben könnte. Aber ausschliessen kann man es nicht.

Alle Firmen versuchen, ihre Kunden an sich zu binden. Besonders beliebt ist das „Vendor-Locking“, bei dem die Daten des Kunden in Geiselhaft für seine Lieferanten-Treue genommen werden. Bis vor kurzem war ein Autor, der sei-

ne Werke mit Microsoft Word schrieb, dieser Firma auf Gedeih und Verderb ausgeliefert. Exporte in andere Formate waren nur möglich, soweit es die Microsoft-Software zuliess und war immer von einigen Verlusten begleitet. Und Microsoft war noch vergleichsweise grosszügig. Wer einmal ein PDF-Formular der Schweizerische Sozialversicherungsanstalt auszufüllen versucht hat, kann ein Lied davon singen, wie wenig grosszügig zum Beispiel die Firma Acrobat mit den Inhalten ihrer Kunden umgeht.

Die schon erwähnte Photo-CD war ein tolles, preisgünstiges, hochqualitatives Scan-Angebot für Kunden. Der Hauptfehler des PCD-Formats ist, dass es niemand mehr richtig entziffern kann, weil seine Spezifikation nie offengelegt wurde. Gerade nach dem Konkurs sitzt der Konkursrichter auf solchen Spezifikationen und Patenten, da er verpflichtet ist, diese nach Möglichkeit im Interesse der Konkursgläubiger zu versilbern. Der einzelne, private Urheber von im PCD-Format gespeicherten Photographien kann solche Forderungen natürlich nicht bezahlen und muss seine Werke einfach verloren geben.

Das E-Book-Format für Amazons Lesegerät Kindle ist geheim. Amazon nimmt sich heraus, gekaufte Bücher auch nach dem Kauf zurückzuziehen. Man kann Daten in diesem Format nicht auf einem anderen Gerät, geschweige denn auf dem Gerät einer anderen Firma nutzen. Dies ist ein klassisches Beispiel, wie ein Format die in diesem Format gespeicherten Daten an proprietäre Software, Hardware und Infrastruktur bindet. Nur Nerds beschäftigen sich mit Fragen, wie man Kindle-Bücher oder iTunes-Songs „befreien“ kann. „Normale“ Benutzer verhelfen hingegen oft den extremsten unfreien Formaten (Word, Amazon-E-Book, ...) zum Durchbruch, da sie sich zuerst einmal nicht für die ferne Zukunft interessieren. Memo-Institutionen dürfen sich deshalb nicht am Durchschnittsnutzer orientieren, denn es ist ihre Pflicht, sich für die fernere Zukunft zu interessieren.

Ein anderes, wichtiges Beispiel für Vendor-Locking sind Oracle-Datenbanken. Diese sind in der Schweiz bei grossen Organisationen extrem weit verbreitet. Oracle hat seinen eigene, nicht standardkonformen Dialekt der Datenbank-Abfragesprache SQL (Standard Query Language) entwickelt. Alle Abfragen und Views, die in dieser Sprache formuliert sind, können nicht auf andere Datenbanksysteme übertragen werden. Sollte Oracle einmal Konkurs machen, wird es sehr schwierig, die Bestände in den Datenbanken der meisten grossen Firmen zu erhalten.

2.4 Abdeckung des Bereichs

Die meisten Dateiformate sind für einen Bereich von Daten zuständig:

- Bildformate dienen zur Speicherung von Bildern;
- Grafikformate dienen zur Speicherung von Vektorgrafik;
- Audioformate zur Speicherung von akustischen Daten;
- Videoformate zur Speicherung von Bewegtbildsequenzen;
- Textformate dienen der Speicherung von Texten;
- Mailformate dienen zur Speicherung von Mail;
- XML-Formate dienen zur Speicherung von partiell strukturierten, hierarchisch organisierten Daten;

- Tabellenkalkulationsformate dienen zur Speicherung von Zahlen, Berechnungen, Formeln und Diagrammen;
- komplexere Dokumentformate dienen zur Speicherung von Texten, Bildern, Tabellen und Diagrammen;
- Datenbankformate dienen zur Speicherung von relationalen Datenbanken.

Diese Aufzählung ist zwar umfassend aber nicht vollständig. Sie dient aber zur Illustration, was mit „Abdeckung des Bereichs“ gemeint ist. Ein brauchbares Bildformat muss das Speichern und adäquate Darstellen eines jeden erdenklichen Bilds ermöglichen. Akustische Daten sind eindimensionale Luftdruckdaten auf einer Zeitachse. Ein für die Langzeiterhaltung taugliches Audioformat sollte jede solche Sequenz im hörbaren Bereich aufzeichnen können. Diese Anforderung kann man so für jeden einzelnen Bereich konkretisieren bis hin zu einem Datenbankformat für relationale Datenbanken, welches jede relationale Datenbank adäquat abbilden können sollte.

Zur Abdeckung des Bereichs gehört auch die Vollständigkeit eines Datenformats. Ein komplexes Dokument ohne eingebettete Schriften verlässt sich darauf, dass diese ausserhalb vorhanden sind. Sind sie dies nicht, so kommt beim Ausdrucken nur ein Zeichensalat heraus, wie man es etwa manchmal bei PDF-Dateien mit exotischen, nicht eingebetteten Schriften erlebt.

2.5 Digitale Qualität

Im Gegensatz zu analogen Daten scheint der Begriff „Original“ für digitale Daten seine Bedeutung weitgehend verloren zu haben. Denn eine bitgetreue „Kopie“ des „Originals“ ist gleichwertig mit diesem und nicht davon zu unterscheiden. Ich kann während dem Schreiben dieses Artikels abwechselnd an verschiedenen „Exemplaren“ des „Originals“ weiterarbeiten, die dann jeweils in der Cloud synchronisiert werden, sodass am Ende keine oder alle das „Original“ darstellen.

Es gibt aber auch im Digitalen „schlechtere“ und „bessere“ Kopien. Die sind natürlich keine bitgetreuen Kopien, sondern Umrechnungen in andere Formate, andere Auflösungen, grössere Kompressionen. Und eine Memo-Institution sollte sich grundsätzlich darum bemühen, ihr Archivgut in der höchstmöglichen Qualität zu speichern. Eine Reduktion dieser Qualität kann später immer noch beim Abrufen als Denormalisierungsschritt vorgenommen werden. Einmal vernichtete Information ist aber unwiederbringlich verloren.

Eine Verschlechterung der Qualität wird aus vielen Gründen vorgenommen. Manchmal wünscht man eine Vereinheitlichung der Grösse und Formate verschiedener Bilder, manchmal möchte man Speicherplatz sparen, manchmal geht es um beschränkte Bandbreite, die ein ruckelfreies Abspielen eines hochqualitativen Videos nicht zulässt.

Die angewendeten Techniken der Verschlechterung sind bei Bild, Ton und Video normalerweise eine Reduktion der Abtastfrequenz, eine Reduktion der Auflösung und eine, die sinnliche Wahrnehmung wenig störende, verlustbehaftete Kompression.

2.5.1 Exkurs Kompression

Kompression wird in vielen Memo-Institutionen als verwerflich angesehen. Dabei wird oft nicht klar unterschieden zwischen reversibler Kompression – etwa in einer ZIP-Datei – und irreversibler Kompression – etwa eines Bildes in einem JPEG-Format oder einer Audiodatei im MP3-Format. Reversible Kompression und alle Formate, welche eine solche beinhalten, ist für die Langzeiterhaltung völlig unproblematisch, solange sie klar standardisiert und nicht mit immaterialgüterrechtlichen Einschränkungen verbunden ist. Heutzutage findet Kompression schon im Gerätetreiber vieler Dateisysteme statt, ohne dass der Benutzer etwas davon merkt. Da bei reversibler Kompression keine Information verloren geht, ist sie äquivalent zu jeder anderen digitalen Kodierung der Daten. Häufig wird noch gegen die reversible Kompression ins Feld geführt, dass sie instabil ist gegen einzelne Bitfehler. Das Problem der nötigen Redundanz als Schutz gegen Bitfehler wird aber in heutigen Informationssystemen anderweitig auf tieferer Ebene gelöst. Eine Kopie oder eine Übermittlung scheitert als Ganzes, wenn sie nicht Bit für Bit korrekt angekommen ist. Redundanz ist allerdings auch für eine Memo-Institution nötig. Um sich gegen Datenverlust zu schützen, sollten alle hochqualitativen Unterlagen in drei bis fünf Kopien an verschiedenen Orten gespeichert sein und diese Kopien laufend automatisch stichprobenartig auf Identität geprüft werden. Nur so kann man defekte Speichermedien frühzeitig detektieren und Datenverlust vermeiden. Ein anderer Ansatz ist natürlich das verteilte, weltweite Archiv, wo alle Inhalte mehrfach bei verschiedenen Memo-Institutionen gespeichert sind. Aber gerade mit dem Siegeszug des Streaming ist eine gefährliche Tendenz – weg von vielen einzelnen individuellen Kopien zur einzigen Kopie weltweit – beobachtbar. Diese Tendenz kann, von übertriebenen Sparbemühungen am falschen Ort (beim Speicherplatz) verstärkt, zu Katastrophen führen.

Bei der irreversiblen Kompression hingegen geht echt Information verloren. Wenn man in mehreren Generationen eine JPEG-“Kopie“ eines Bildes von der vorhergehenden JPEG-“Kopie“ generiert, nimmt die Qualität des wiedergegebenen Bildes sukzessive ab. Dasselbe gilt für MP3-“Kopien“ von Audio-Daten und für Flash-“Kopien“ von Video-Daten. Aus diesem Grund eignen sich irreversibel komprimierende Formate nicht für die Langezeitarchivierung. Dass deshalb grössere Datenmengen archiviert werden müssen, als es mit maximaler irreversibler Kompression nötig wäre, ist kein Argument für die Archivierung schlechter Kopien. Verglichen mit dem Aufwand für die Langzeiterhaltung überhaupt, fällt der Preis der Speichermedien nicht ins Gewicht.

2.5.2 Authentizität bei digital entstandenem Archivgut und bei der Digitalisierung

Selbstverständlich kann die Qualität bei digital entstandenem Archivgut nicht höher sein, als das ursprüngliche Werk. Wenn eine Foto schon von der Kamera als JPEG abgespeichert wurde, kann niemand die Format-Artefakte entfernen, weil es kein originaleres Original gibt, das durch die irreversible Kompression verfälscht wurde. Jedes Bildformat enthält aber Angaben darüber, welche Farbe an welchem Fleck auf dem Bild zu sehen ist. Diese Angaben können auch von Daten in nicht archivtauglichen Formaten getreu in einem Normalisierungsschritt in standardisierte, höchstens reversibel kompri-

mierende Formate – z.B. Das PNG-Format – konvertiert werden, ohne dass dabei weitere Information verloren geht. Die im Magazin des Langzeitarchivs aufbewahrte normalisierte Version ist maximal authentisch, wenn sie gegenüber dem eingelieferten Original keine – oder möglichst wenig – Information verloren hat. Idealerweise ist die Normalisierung reversibel. D.h. im Beispiel der JPEG-Datei sollte die Denormalisierung aus dem PNG-Format in eine JPEG-Datei gleicher Qualität wieder die Originaldatei ergeben. Eine Notwendigkeit, nicht vorhandene Auflösung mit künstlichem Aufblasen der Daten (etwa: Skalieren und Interpolieren der Bilder) zu simulieren, besteht nicht. Vielmehr sollten etwa Digitalisate alter Fotos oder Schallplatten nicht deren Defekte „reparieren“. Die Reparatur kann ruhig den Benutzern oder dem Denormalisierungsschritt überlassen werden, denn sie trägt nichts zur im Datensatz enthaltenen Information bei.

Wenn das Archivgut mittels Digitalisierung analoger Daten entsteht, stellt sich die Qualitätsfrage anders: Wie hoch soll die Auflösung in der Ebene und die Farbauflösung sein, damit von einer authentischen Abbildung gesprochen werden darf? Dabei darf man sich keiner Illusion hingeben: Bei der Digitalisierung geht immer etwas verloren. Das Papier oder die Schellackplatte kann nicht mehr chemisch untersucht werden. Das Materielle muss anderweitig für die Erhaltung dokumentiert werden. Solange es aber um von menschlichen Sinnesorganen wahrgenommene immaterielle Inhalte geht, kann man die Qualität so wählen, dass die Grenzen der menschlichen Wahrnehmung erreicht werden. Ein Bildformat kann heute üblicherweise pro Bildpunkt 16'777'216 Farbwerte wiedergeben. Das menschliche Auge kann höchstens 4'000 bis 10'000 Farben unterscheiden. Und auch das nur, wenn eine grössere Anzahl von Bildpunkten in der Umgebung dieselbe oder eine ähnliche Farbe hat. Es besteht also kein Anlass, eine noch grössere Farbauflösung zu wählen. Sehr wesentlich ist aber natürlich eine klare Spezifikation und Dokumentation des zugrundeliegenden Farbraums, des beim Digitalisieren eingesetzten Lichts etc., damit die Farbechtheit gewährleistet werden kann. Auch die geometrische Auflösung des Auges ist beschränkt, wie alle wissen, die einmal beim Augenarzt die Buchstabentafel vorgelesen haben. Es scheint sinnlos, die Auflösung des Digitalisats massiv grösser anzusetzen als was man mit dem Auge unterscheiden kann. Auch die physikalische Grösse des Originals sollte dabei respektiert werden. Das alles gilt natürlich nur für Bilder, deren Hauptzweck es ist, von Menschen angeschaut zu werden. Eine wissenschaftliche Messung – etwa ein Satellitenbild – kann sehr viel wichtige und nützliche Information jenseits der menschlichen Wahrnehmungsgrenze enthalten. In diesem Fall ist die Auflösung der eingesetzten Messinstrumente und nicht die des menschlichen Auges massgebend. Analog gibt es eine untere und obere Grenze in Lautstärke und Frequenz wo das menschliche Ohr kein Geräusch mehr wahrnimmt. Die Auflösung der Abtastung der Luftdruckschwankungen am Mikrophon braucht nicht viel höher zu sein. Damit ist die ideale akustische Auflösung nicht weit von derjenigen der klassischen CD-Audio entfernt. Bei Videodaten, die aus Bild- und Toninformation zusammengesetzt sind, gelten ähnliche Überlegungen.

2.5.3 Qualität von Textdokumenten

Bei komplexeren Textdokumenten stellt sich die Qualitätsfrage noch einmal anders. Schrift ist ein Code zur Darstellung akustischer Sprachsignale. Ein guter – wenn auch nicht abschliessender – Test für die Qualität von Textformaten besteht darin, sie von einem Sprechprogramm vorlesen zu lassen. Dabei stellt sich schnell heraus, dass Texte in Bilddateien von tiefster Qualität sind, denn die Sprache kann nicht aus den Pixeln rekonstruiert werden. Diese geben nur wieder, an welchen Stellen die Seite schwarz oder weiss ist, bzw. welche Farbe dort sichtbar ist. Auch Seitenbeschreibungssprachen wie das auf PostScript beruhende PDF-Format von Acrobat sind nicht viel besser. Diese geben wieder, welche Buchstaben, in welcher Grösse und Schrift, wo auf der Seite zu finden sind. Aber sie enthalten keinerlei Angaben zur Textlogik. Ein Vorleseprogramm wird bei doppelspaltigem Satz oder bei Tabellen hoffnungslos scheitern. Suchalgorithmen finden Wörter nicht, wenn sie am Zeilen-, Spalten- oder Seitenende getrennt wurden. Denn die Aufteilung in visuelle Zeilen, Spalten und Seiten ist nur ein Hilfskonstrukt, um die Sprachsignale visuell wiederzugeben. Für das Verständnis des Inhalts tragen sie aber im Allgemeinen nichts bei oder sind sogar sinnstörend.

Memo-Institutionen, die ganze Bücher als hochaufgelöste Bilddateien speichern, liefern also tiefe Qualität, denn nicht die Schriftart, sondern der Text ist das Wesentliche. Nur Textdokumente, die barrierenfrei vorlesbar sind, haben hohe Qualität. Was für Behinderte gut ist, ist auch für die Erhaltung des kulturellen Erbes gut. Qualität von Textdokumenten misst sich auch an der Möglichkeit, korrekt und vollständig nach Textstellen zu suchen oder einen zusammenhängenden Absatz zu markieren und zu kopieren, der über Zeilenumbrüche, Spaltenumbrüche und Seitenumbrüche hinausgeht. Der ultimative Test der Qualität eines Textdokuments besteht im Versuch, das Layout oder die Schrift zu ändern. Nur wenn dies – weitgehend, d.h. bis auf Platzierung von Illustrationen etc. – erfolgreich ist, kann von hoher Qualität eines Textdokuments die Rede sein.

2.6 Format-Identifikation

Ein Dateiformat, selbst eines, das sich für die Langzeiterhaltung eignet, wird nicht auf ewig unverändert bleiben. Wenn neue Erkenntnisse, neue Möglichkeiten und neue Ausdrucksformen zu neuen Formaten oder neuen Versionen von Formaten Anlass geben, sollte dies für das Format lesende, interpretierende, darstellende oder denormalisierende Programme erkennbar sein. Bei den oben erwähnten einfachen Textdateien fehlt etwa eine solche Identifikation, die es erlaubt, zweifelsfrei zwischen ASCII, EBCDIC, ISO Latin und UTF-8 zu unterscheiden.

Glücklicherweise ist es bei fast allen heute gebräuchlichen, für die Langzeiterhaltung brauchbaren Formaten der Fall, dass das Format und seine Version in jeder dem Format entsprechenden Datei gespeichert sind. Das JHOVE-Projekt hat sich zum Ziel gesetzt, jeder denkbaren Datei ein eindeutiges Format zuzuordnen. Schwieriger wird dies bei binärem oder Quellcode und bei Messdatenreihen. In diesen Bereichen hat sich die klare Format-Identifikation noch nicht überall durchgesetzt.

2.7 Eingebettete Metadaten

Metadaten sind normalerweise stärker strukturiert als das digitale Archivgut. Sie enthalten Angaben zum Kontext der Inhalte. Wenn sie verändert werden, ändert das nichts an der Authentizität der Primärdaten, sondern höchstens an deren Interpretation. Nach der initialen Normalisierung dürfen Primärdaten im Archiv nie mehr verändert werden. Metadaten hingegen schon.

Metadaten enthalten also üblicherweise Angaben zum Titel, zum Urheber, zum Ort, zur Zeit, zur Quelle, zum rechtlichen Status der archivierten Inhalte. In typischen Archiven heute sind sie in der Datenbank der Memo-Institution gespeichert, welche als Katalog dient. Es ist aber sehr wichtig, dass solche Metadaten in das Archivgut eingebettet werden und sich auch bei der Denormalisierung mit den Inhalten verbreiten. So wird die Rechtssicherheit der Benutzer erhöht, welche schon der Datei entnehmen können, was ihre Rechte sind. Verteilte Archive werden ermöglicht, weil der Katalog aus den eingebetteten Metadaten des Archivguts aufgebaut werden kann. Die heutige Unart der meisten Memo-Institutionen, ihre Digitalisate nur unter strengstem Verbot jeglicher kommerzieller Nutzung zugänglich zu machen, und damit jegliche wissenschaftliche Beschäftigung mit den Inhalten zu verhindern, würde überflüssig. Denn meistens geht es diesen Institutionen gar nicht darum, illegalerweise die Rechte der Allgemeinheit an gemeinfreien Werken zu beschneiden, sondern sie wollen als Quelle sichtbar sein und genannt werden, weil davon ihr Budget in der Verwaltung und ihr Ruf in der Bevölkerung abhängt.

Erfreulicherweise ermöglichen die meisten Formate heute das Einbetten von Metadaten. Mit dieser Einschränkung schliesst man also kaum wichtige Formate aus. Reine Textdateien erfüllen diese Anforderung aber nicht.

3 Digitale Inhalte, die wir (noch) nicht archivieren können

Aufgrund der Anforderungen an Normalformate für digitale Inhalte, können gewisse digitale Inhalte vorläufig nicht auf Dauer archiviert werden. Das betrifft in erster Linie Inhalte, die so innig mit proprietären Hardware- oder Software-Produkten einzelner Firmen verbunden sind, dass die Wahrscheinlichkeit sehr klein ist, dass sie in fünfzig Jahren oder mehr adäquat genutzt oder verstanden werden können.

3.1 Verschlüsselte Inhalte

Viele Formate unterstützten die Verschlüsselung einzelner Teile der Inhalte. Alle Dateien können als Ganzes verschlüsselt werden. Auf der anderen Seite kommt Kryptographie bei der Signatur zum Einsatz.

Es scheint – vorläufig – unrealistisch, dass eine konsistente Schlüsselverwaltung im Archiv über fünfzig Jahre durchgeführt werden kann. Ausserdem verhindert Verschlüsselung den freien Austausch verteilter Inhalte zwischen verschiedenen Archivstandorten. Deshalb werden grundsätzlich keine verschlüsselten Inhalte archiviert.

Eine kryptographische Signatur von Inhalten, stört jedoch nicht. Nach Verlust der relevanten Schlüssel kann schlimmstenfalls die Signatur nicht mehr überprüft werden. Die Authentizität der Primärdaten wird dadurch nicht verändert.

Selbstverständlich kann ein Archiv sämtliche Archivinhalte verschlüsseln oder auf verschlüsselten Laufwerken ablegen, um unbefugten Zugriff auf die Unterlagen zu verhindern. Dabei werden archiveigene Schlüssel verwendet. Aus der Sicht der Archivstruktur ist das eine „reversible Speicherung“ wie die reversible Kompression, solange die Entschlüsselung jederzeit möglich ist. Solche „externe“ Verschlüsselung ist also mit dem Verbot von Verschlüsselung der Inhalte nicht gemeint.

3.2 Binärer Programmcode

Binärer Programmcode ist normalerweise nur im Kontext eines speziellen Betriebssystems lauffähig. Betriebssysteme werden weiterentwickelt und Firmen verschwinden. Schon heute kommt es oft vor, dass binärer Code für eine ältere Version eines Betriebssystems auf der neueren nicht mehr lauffähig ist. Das liegt *nicht* daran, dass er „binär“ ist oder seine Struktur nicht genügend offen standardisiert ist. Sondern man muss über einen Zeitraum von fünfzig Jahren davon ausgehen, dass heutige Recheneinheiten, Eingabegeräte etc. nur noch dunkel erinnert werden und von zukünftigen Geräten nicht mehr unterstützt werden. Somit dürfte praktisch jeder binäre Programmcode nicht mehr lauffähig sein, weil er am Vendor-Locking (Abhängigkeiten zu obsoleter Hardware, Betriebssystem, Sicherheitskonzept) scheitert.

Vorläufig wird in CultLib kein binärer Programmcode archiviert.

3.3 Quellcode

Quellcode ist in verschiedenen Programmiersprachen schön standardisiert. Allerdings ist die Abhängigkeit von Modulen im selben Projekt so eng, dass

nur der gesamte kompilierbare Quellcode eines Programms in Gänze verständlich bleibt. Ausserdem hängen die meisten Programmiersprachen und somit die Programme enger mit dem Betriebssystem und proprietären Geräten (Maus, Bedienoberfläche) und Konzepten (Filesystem) zusammen als man auf den ersten Blick glauben möchte. Quellcode kann also eigentlich aus den selben Gründen nicht über grosse Zeiträume erhalten werden, wie binärer Programmcode. Eine Ausnahme bilden Programme, die nur mit sehr wenigen externen Betriebssystem-Routinen kommunizieren. Etwa Konversionen mit einem endlichen Eingabe- und einem Ausgabe-Byte-Strom. Der Quellcode solcher Programme könnte theoretisch erhalten werden. Es ist aber davon abzuraten, ausser er werde von einem Tangle- & Weave-Mechanismus wie Donald Knuths „TEX: The Program“ in ein Dokument, ein Buch, verwandelt, das sich an menschliche Leser richtet.

Vorläufig wird in CultLib kein Quellcode archiviert. Relativ einfache Formeln aus einem klar definierten Satz von Funktionen (etwa in Excel-Dateien) gelten dabei nicht als „Programmcode“. Kleine Stücke von VB-Script können als Teil eines Office-Dokuments archiviert werden, wenn sie auch in fünfzig Jahren kaum mehr funkionieren werden, aber eventuell zum Verständnis beitragen.

3.4 Digitale Kunst

Viele moderne „digitale Kunstwerke“ sind Installationen, die extrem abhängig von Eingabe- und Ausgabe-Geräten und proprietärer Hard- und Software sind. Der „digitale“ Reiz liegt nämlich oft in einer Interaktivität, welche herkömmlichen Dokumenten nicht innewohnt. Es handelt sich eigentlich um spezielle Programme mit Kunstanspruch. Damit ist ihre Erhaltung ausschliesslich auf Emulationen angewiesen und völlig illusorisch, wenn sie auf Firmengeheimnissen basieren, die nicht emuliert werden können.

Vorläufig werden in CultLib keine „digitalen Kunstwerke“ archiviert.

3.5 Games

Games sind sozusagen eine Form von „digitaler Kunst“, wo die Interaktion im Zentrum steht. Die Argumente gegen die Langzeiterhaltung digitaler Kunst treffen auch auf Games zu. Ausserdem werden ihre Innereien äusserst proprietär geheimgehalten, was der Möglichkeit einer erfolgreichen Archivierung jeden Boden entzieht. Am Beispiel von modernen Emulationen von Pac-Man auf modernen Rechnern sieht man, dass höchstens Emulationen als Basis für die Archivierung dienen könnten.

Vorläufig werden in CultLib keine Games archiviert.

3.6 Websites

Websites sehen auf den ersten Blick recht archivierbar aus. Sie bestehen aus gut standardisierten Teildateien in einem Verzeichnisbaum in wenigen klar definierten Formaten (HTML, CSS, ...). Das gilt allerdings nur für statische Websites, wo serverseitig keine „Webanwendung“ mit viel Code irgendwelche Daten aus Datenbanken zu dynamisch erzeugte HTML-Dateien zusammenfügt. Ausserdem gibt es kaum eine Webseite, die nicht zusätzlich zu CSS auch intensiven Gebrauch von clientseitiger Programmierung in JavaScript

oder einer anderen Skriptsprache macht. Schliesslich ist die Abgrenzung, was „intern“ und was „extern“ ist, recht schwierig. Aufgrund des Programmcode-Verbots wären Websites also nur in den uninteressantesten Fällen archivierbar. Aus diesem Grund ist wohl eher auf eine solche Archivierung zu verzichten. Es kann nützlicher sein, stattdessen die zugrundeliegende Datenbank einer Webanwendung zu archivieren.

Vorläufig werden in CultLib keine Websites archiviert. Aber Websites werden benutzt, um CultLib-Inhalte zugänglich zu machen.

3.7 Zeichenfolgen

Reiner Text ist eine Datei, deren Bytes als Zeichenfolge in einem Zeichensatz (ASCII, EBCDIC, ISO Latin, UTF-8) interpretiert werden, welcher angibt, welches Zeichen für welchen Buchstaben steht. Solche Textdateien haben sich zwar bisher als sehr haltbar erwiesen. Es bestehen aber einerseits ernsthafte Probleme der Identifikation des verwendeten Zeichensatzes und andererseits werden keine „eingebettete“ Metadaten unterstützt, die klar von Primärdaten getrennt sind.

XML gilt in diesem Zusammenhang nicht als blosse Zeichenfolge. XML-Dateien können archivtauglich sein, wenn sie eine freie und offene standardisierte Formatbeschreibung (XSD, DTD) referenzieren, welche den obigen Anforderungen genügt.

In CultLib werden keine reinen Texte archiviert. Stattdessen ist für solche Inhalte ein Dokument-Format als Behälter zu wählen.

4 Archivtaugliche Dateiformate

Auf der Basis der obigen Überlegungen kommen wir zur Aufzählung der wichtigsten in CultLib unterstützten Bereiche und der Dateiformate, welche uns dafür geeignet scheinen. Die vielen Abkürzungen für Formate werden hier nicht weiter erklärt. Stattdessen verweisen wir auf die entsprechenden Wikipedia-Artikel für vertiefte Information.

4.1 Bild-Dateien

Ein Bild ist vollständig gegeben, wenn für jeden Bildpunkt klar ist, welche Farbe er hat. Damit eignen sich praktisch alle standardisierten Bildformate (ohne proprietäre Erweiterungen, wie sie etwa oft beim TIFF-Format angetroffen werden) für die Langzeiterhaltung.

Das sehr populäre JPEG-Format geht allerdings wie oben beschrieben meistens mit einer irreversiblen Kompression einher. Das JPEG2000-Format ermöglicht zwar die Vermeidung solcher irreversibler Kompression, ist aber immer noch unnötig kompliziert und vierzehn Jahre nach seiner Konzeption immer noch weniger verbreitet. Zum Zeitpunkt der Standardisierung konnten sich die industriellen Beteiligten nicht auf einfache Formate einigen. Ausserdem war die Ökonomie des Speicherplatzes („Bits-Sparen“) noch oberstes Gebot. Das führte dazu, dass ein heilloses Sammelsurium von Formaten unter dem Namen „JPEG“ (Joint Picture Expert Group) zusammengefasst wurde, dessen viele Teilformate praktisch von keiner Software vollständig standardgemäss behandelt werden.

Aus diesem Grund haben wir für CultLib das freie und offene PNG-Format als Normalisierungsformat festgelegt. Dieses beschreibt die Pixelfarben genau, enthält Möglichkeiten für 1 Bit/Pixel (schwarz/weiss), 1 Byte/Pixel (Graustufen), 1 Byte/Pixel (Palette) und 3 bzw. 4 Byte/Pixel (Farben und Transparenz-Kanal). Das PNG-Format unterstützt die Einbettung von Metadaten. Jedes Bild kann in einem Normalisierungsschritt adäquat ohne Informationsverlust in das PNG-Format konvertiert werden.

Als Denormalisierungen werden Konversionen ins JPEG-Format und Skalierungen angeboten.

4.2 Audio-Dateien

Tondaten sind vollständig gegeben, wenn die Luftdruckschwankungen an einem (oder mehreren) Orten als Funktion der Zeit im hörbaren Bereich rekonstruiert werden können. Damit eignen sich fast alle standardisierten Audio-Formate für die Langzeitarchivierung.

Das WAV-Format, welches dem CD-Audio-Format sehr nahe ist, ermöglicht allerdings keine gut standardisierte Einbettung von Metadaten. Das MP3-Format ist zwar sehr verbreitet, führt aber immer zu verlustbehafteter Kompression und ist von der Fraunhofer-Gesellschaft mit Patenten belegt und somit nicht frei und offen verfügbar. Das Vorbis (Ogg)-Format ist eine freie Alternative zum MP3-Format, führt aber ebenfalls zu verlustbehafteter Kompression.

Aus diesen Gründen haben wir für CultLib das FLAC-Format als Normalisierungsformat festgelegt. Jede Audio-Datei kann in einem Normalisierungs-

schritt adäquat ohne Informationsverlust in das FLAC-Format konvertiert werden.

Als Denormalisierung wird das MP3-Format zum Download oder gestreamt angeboten.

4.3 Video-Dateien

Das Gebiet der Videodaten war schon in analogen Zeiten hart umkämpft und von inkompatiblen Standards dominiert wie von den Fernseh-/Video-Formaten PAL (Europa), NTSC (USA) und SECAM (Frankreich) und den Film-Formaten (8mm, 16mm, 32mm Film; 15, 24, 25, 30 Bilder pro Sekunde; quadratisch, 4:3, 9:6, Breitleinwand, 3D usw.; mit Lichtton, Magnetton ...).

Das liegt schon daran, dass weit weniger klar ist, welches der Bereich ist, der von einem Bewegtbild-Format abgedeckt werden sollte. Auf den ersten Blick würde man meinen, dass sich Videodaten aus Bilddaten und Tondaten als Funktion der Zeit zusammensetzen. Sie wären also vollständig gegeben, wenn die Farbe und der Luftdruck (evtl. an mehreren Orten) an jedem Bild- und Zeitpunkt eindeutig gegeben sind. Wegen der notwendigen Synchronisation kommen noch Anforderungen an Timestamps hinzu. Diese bilden auch die Basis für allfällige Untertitel eventuell in vielen Sprachen. Untertitel werden oft ignoriert und unterschätzt. Sie sind aber ein wesentliches Instrument nicht nur für Fremdsprachige und Behinderte, sondern für die Durchsuchbarkeit, Lokalisierbarkeit und Analysierbarkeit bewegter Bildinhalte. Und manche möchten auch 3-D-Filme im selben Format speichern können. Für letztere wird allerdings wohl besser ein anderes, separates Fileformat eingesetzt.

Im der heutigen digitalen Welt findet eine förmliche Standardisierungs- und Patentierungsschlacht der Formate für Videodaten statt, wo sich Betriebssysteme, Browser- und Mediaplayer-Hersteller bekämpfen und der Nutzer auf Schritt und Tritt dafür bestraft wird, dass er das falsche Betriebssystem, die falsche Umgebung benutzt. Weiter verschärft wird diese Unklarheit für den Normalbenutzer, dass viele Formatbezeichnungen – wie etwa MP4 – eigentlich nur eine Standardisierung des Behälters festlegen, aber die Bild-, Audio, und Untertitel-Inhalte – je nach verwendetem „Codec“ – verschiedene Formate haben können. Am weitesten verbreitet und somit aktuell wohl am stabilsten dürfte das Format der DVD sein. Die verschiedenen High-Definition-Formate sind noch nicht konsolidiert. Es ist auch unklar, ob dafür eine genügend starke Nachfrage besteht.

Da Videodaten naturgemäss noch viel mehr Speicherplatz und Bandbreite beanspruchen als Audio- oder Bilddaten – ganz grob sind Audiodaten eindimensional, Bilddaten zweidimensional und Videodaten dreidimensional –, werden auf diesem Gebiet auch sehr viele irreversible Kompressionstechniken eingesetzt. Ein verlustloses Format ist momentan kaum in Sicht. Ausserdem ist der Kampf um Streaming-Strategien sehr viel pointierter als im Audiobereich. Auch sind die Anforderungen an die statische Farbauflösung eines Pixels beim bewegten Bild kleiner. Meistens wird die Pixelfarbe mit nur 8 Bit beschrieben.

Es ist relativ einfach, eine Wunschliste für das ideale Videoformat zusammenzustellen:

- freies Containerformat, das Videodaten, mehrere Audiodaten und mehrere Untertiteldaten enthalten kann,
- freies Videodatenformat mit Timestamps für Synchronisation,
- freies Audiodatenformat mit Timestamps für Synchronisation,
- freies Untertitelformat, das mit Timestamps korrelierte Textdaten enthält,
- keine irreversible Kompression oder zumindest im Audiobereich nahezu verlustlose Kodierung,
- klar definierbarer Farbraum für die Videodaten,
- frei wählbare Auflösung in Zeit und Raum.

Selbstverständlich sollten alle beteiligten Formate die Anforderungen an für die Langzeiterhaltung geeignete Normalformate erfüllen.

Anhand dieses Wunschkatalogs kann man leicht beschreiben, warum gewisse Formate ungeeignet sind für die Langzeiterhaltung:

- *Flash:* Proprietär (Acrobat), auf Streaming ausgerichtet, FLV (H.263-ähnlich), MP3, keine hochqualitativen Audiodaten, oft Probleme mit der Aspect Ratio (4:3 oder 16:9).
- *MOV (Quicktime):* Proprietär (Apple), vor allem Container, gleiche Codecs wie MPEG-4, war Basis für MPEG-4
- *AVI:* Proprietär (Microsoft), relativ alt und unflexibel
- *WMV:* Proprietär (Microsoft), DRM, proprietäre Codecs, oft nicht abspielbar auf Nicht-Microsoft-Plattformen
- *MPEG-4:* relativ offener ISO-Standard, generelles Format für zeitbasierte Inhalte (Medienbasisformat), oft proprietäre Codecs, DivX, XviD, H.263, H.264 (AVC) H.265, HEVC (Patente bei MPEG LA); HE-AAC, Apple als Registrations-Autorität für Codecs, spezielle DRM-Unterstützung
- *MPEG-2:* relativ offener ISO-Standard, H.222, H.262, verbreitetes DVD-Format, von Profis genutzt, relativ hochauflösend
- *WebM:* freie und offene von Google als Alternative zu Flash und MPEG-4 entwickelter Standard, auf Streaming ausgerichtet, offenes Containerformat Matroska mit freiem VP8 oder VP9 als Videoformat und ebenso freiem Vorbis (Ogg) oder opus als Audioformat, WebVTT als Untertitelformat. von vielen Browsern im HTML5 <video>-Tag auf vielen Plattformen unterstützt. Sowohl Video- als auch Audio-Daten werden nicht verlustlos komprimiert.

Obwohl vor allem die neueren dieser Formate vor allem im Low-Endbereich auf YouTube für Streaming eingesetzt werden, lassen die meisten auch hochaufgelöste, professionelle Inhalte zu. Völlig verlustlose Kompression scheint es aktuell bei keinem der verbreiteten Video-Formate zu geben.

Für CultLib werden Videodaten in einem Normalisierungsschritt in WebM-Dateien verwandelt. Die dazu eingesetzte Software stammt nach Möglichkeit aus dem WebM-Projekt selber oder aus dem FFmpeg-Projekt.

4.3.1 Exkurs DRM

Unter Digital Rights Management (DRM), auch Technische Schutzmassnahmen (TSM) genannt, versteht man ein Bündel von oft mit kryptographischen Methoden arbeitenden Techniken, die dazu dienen sollen, unautorisierte Kopien und Nutzungen von urheberrechtlich geschützten Inhalten zu verhindern. Die meisten dieser Ansätze benötigen beim Abspielen einen online-Zugriff auf Server der Rechteinhaber, wo die Rechtmässigkeit der abgespielten oder kopierten Kopie geprüft wird. DRM haben sich für ehrliche Käufer als grosses Problem erwiesen, während sie von Piraten relativ leicht ausgehebelt wurden. Aus diesem Grund hat das schweizerische Institut für geistiges Eigentum eine Beobachtungs- und Meldestelle für Probleme mit technischen Schutzmassnahmen eingerichtet. Nachdem mehrere Tausend ehrliche Käufer ihre Songs nicht mehr abspielen konnten, weil der Server der Rechteinhaberin nicht mehr verfügbar war, sank das Vertrauen der Allgemeinheit in DRM-geschützte Inhalte. Heutzutage werden Songs mehrheitlich wieder ohne DRM verkauft, während E-Books (Amazon) oft mit DRM-Restriktionen verbunden sind und zum Teil vom Verkäufer nach dem Kauf auf dem Gerät des Kunden gelöscht oder zurückgezogen werden, ohne dass eine freie Kopie angelegt werden konnte.

Für die Langzeiterhaltung ist jegliche Form von DRM Gift. Es ist absolut unrealistisch, ein DRM-System fünfzig Jahre lang aufrechterhalten zu wollen. Es sind nur DRM-freie Inhalte archivierbar. Die Formate, die spezielle DRM-Unterstützung enthalten können, sind mit grösster Vorsicht zu geniessen.

Da in CultLib ohnehin nur frei zugängliche Inhalte archiviert werden, sind diese grundsätzlich DRM-frei.

4.4 OOXML/A-Dateien

Sehr viele wichtige Unterlagen entstehen in sogenannten Office-Programmen. Hier hat sich Microsoft Office seit Jahrzehnten weltweit durchgesetzt und wird heute nur von der OpenOffice/LibreOffice-Initiative nennenswert konkurrenziert. Entsprechend sind die drei Officedatei-Formate für Dokumente, Tabellenkalkulation und Präsentationen sehr weit verbreitet. Diese eigneten sich früher ausnehmend schlecht für die Langzeiterhaltung, weil ihre internen Formatspezifikationen nicht nur proprietär im Eigentum von Microsoft waren, sondern auch geheimgehalten wurden.

Die Situation hat sich mit der Einführung der Anwendung Microsoft Office 2007 und ihren Nachfolgern Office 2010, Office 2012 grundlegend verändert. Neu hat die Firma Microsoft die Speicherformate ihrer Programme offengelegt und sogar als ISO-Standards publiziert. Die neuen Formate sind gemeinschaftlich unter dem Namen OOXML (Office Open XML) bekannt geworden. Es handelt sich um Word- (.docx), Excel- (.xlsx) und Powerpoint-Dateien (.pptx). Parallel dazu definierten die Entwickler von OpenOffice, der Open-Source-Konkurrenz von Microsoft Office, ihre Formate für Writer (entspricht Word), Calc (entspricht Excel) und Impress (entspricht Powerpoint) unter dem Namen ODF (Open Document Format), welches ebenfalls als ISO-Standard vorliegt.

Sowohl OOXML als auch ODF speichern die Inhalte eines Dokuments als ZIP-Datei, welche die eigentlichen Dokumente als Kollektion von XML-Dateien

und binären (z.B. Bilddateien) Daten enthält. Beide Formate enthalten aber Komponenten, die nicht für die Langzeitarchivierung tauglich sind: Referenzen auf externe Zeichensätze, Programmcode, Einbettung fremder, binärer Inhalte (z.B. Multimedia-Daten), die von anderen Anwendungen dargestellt werden müssen, und deren Struktur im Normalfall nicht archivtauglich ist. Da die Mehrheit der Dokumente keine solche untauglichen Elemente enthält, scheint es wünschenswert, für die Langzeitarchivierung taugliche Teilmengen dieser Formate zu definieren – analog wie der ISO-Standard PDF/A als Teilmenge des Industriestandards PDF 1.4 definiert wurde.

Da das OpenOffice/LibreOffice-Paket zum heutigen Zeitpunkt in der Allgemeinheit noch keine sehr grosse Verbreitung gefunden hat, und – mindestens bis zur Version 3.x – keine Einbettung der verwendeten Zeichensätze ermöglicht, sind vorläufig nur archivtaugliche Teilmengen der Microsoft-Formate DOCX/A, XLSX/A und PPTX/A möglich.

Auf der Basis des OOXML-ISO-Standards hat die Enter AG eine Teilmenge OOXML/A ausgesondert, welcher viele Office-Dokumente genügen.

In CultLib werden OOXML/A-konforme Officedateien archiviert, deren Komponenten alle in archivtauglichen Formaten vorliegen. Dies sind typischerweise typographisch gestaltete Texte (inkl. Tabellen, Listen, ...) mit Bildern, einfachen Grafiken und Diagrammen als Illustrationen. Spezielle Teildokumente wie mathematische Formeln oder Musiknoten sind damit nicht abgedeckt. Für formelintensive Dokumente kann später das TEX-Format homologisiert werden. Für Musiknoten sind ganz eigene Formate nötig.

Die Grundsätze, die für die Archivtauglichkeit der OOXML/A-Teilmenge angewendet wurden, sind dem ISO-Standard 19005-1 für PDF/A entnommen. Sie betreffen:

- Dateistruktur (Ausschluss von nicht archivtauglichen Inhalten),
- obligatorisches Einbetten notwendiger Inhalte (Bilder, Schriften),
- Ausschluss von Multimedia-Inhalten (Ton, Video),
- Ausschluss von Programmcode,
- Ausschluss von Verschlüsselung,
- obligatorische korrekte Metadaten.

4.5 Relationale Datenbanken

Relationale Datenbanken sind nur sehr schwierig zu archivieren. Die Hersteller haben sich auf einen relativ alten Standard der Abfragesprache SQL geeinigt. Dieser wird allerdings von den meisten nicht eingehalten bzw. proprietär erweitert. Immerhin hat es sich als möglich erwiesen, die eigentlichen Tabellendaten in den Basistabellen auf kontrollierte Weise in XML-Dateien zu konvertieren, die in einer ZIP-Datei zusammengefasst sind.

Dieses Speicherformat heisst SIARD-Format (Software-Independent Archival of Relational Databases) und wurde von der Enter AG für das schweizerische Bundesarchiv entwickelt. Heute hat es den Status eines eCH-Standards (eCH-0165 v1.0). Ausführlichere Beschreibungen findet man auf der einschlägigen Seite des Bundesarchivs und im Artikel der Enter AG über die Archivtauglichkeit des Formats.

In CultLib werden wohl in nächster Zukunft keine Datenbanken archiviert. Wenn sich dies aber aufdrängen sollte, wird dafür das SIARD-Format eingesetzt.