

Werk und Werkausprägung

Langzeiterhaltung digitaler Inhalte für das CultLib-Projekt III

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

CultLib-ID: 88e836cf-c830-44e9-9e60-0a08e0aba500

Version: 1.0

Datum: 14. April 2014

Verfasser: Dr. sc. math. Hartwig Thomas



<http://creativecommons.org/licenses/by-sa/3.0/ch/>

Übersicht

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

Die Berichte über Aspekte der Langzeiterhaltung¹ 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 einem verteilten Repository müssen die Werke und die Werkausprägungen unabhängig von der physischen oder virtuellen Lokalisierung des Archivs eindeutig identifiziert werden.

Für die Identifikation von Werken wird die Verwendung eines Universal Unique Identifiers (UUID) in den eingebetteten Metadaten vorgesehen. Damit ist keine Registrierung bei einer zentralen Verwaltung von IDs nötig. Für die Identifikation von Werkausprägungen (konkreten Repräsentationen des Werks) dient ein „Message Digest“ über die Primärdaten. Dieser kann automatisch aus den Primärdaten berechnet werden.

Werk-UUID, Primärdaten-Digest, Lizenzbestimmungen und Publikationsdatum können den legalen Zustand eindeutig definieren.

Die URLs von CultLib-Dokumenten sind abgeleitet von Domäne, Werk-UUID und Primärdaten-Digest. Jede URL ist permanent in dem Sinne, dass dieselbe URL nie für ein anderes Objekt steht. Sofern sie überhaupt aufgelöst wird, führt sie auf das gewünschte Objekt.

Anforderungen von Listen oder von denormalisierten Objekten werden für das verteilte Repository nicht streng reguliert. Vorschläge werden exemplarisch angeführt.

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 Verteilte Repositorien.....	4
2 Werke und ihre Ausprägung.....	6
2.1 Werkausprägung und Werkexemplar.....	6
2.2 Ausprägung und Abwandlung.....	6
2.3 Sind Aufführungen bloss Ausprägungen?.....	7
2.4 Sind Sammlungen auch Werke?.....	7
3 Identifikation von Werk und Werkausprägung.....	8
4 Generelle Pflicht zur Identifikation?.....	10
5 URL-Struktur für die Anforderung von Dokumenten.....	11
5.1 URLs mit Abfragen.....	12
5.2 URLs und Metadaten.....	12
5.3 URLs und andere Listen.....	12

1 Verteilte Repositorien

In Zukunft werden wohl auch die digitalen Memo-Institutionen einen weltweiten Verbund bilden wie dies ansatzweise heute die Bibliotheken sind. Die verschiedenen Repositorien würden dann teilweise überlappende Bestände lokal halten und lokal nicht vorhandene Werke im Verbund anfordern. Ein Benutzer könnte bei seinem lokalen Archiv² jedes Werk des weltweiten Verbunds bestellen. Die dezentrale Replikation der Werke an verschiedenen Orten ist schon darum erwünscht, weil damit das Risiko des Totalverlusts von Werken minimiert wird, auch wenn ein Archiv wegen einer lokalen Katastrophe völlig ausfällt und seine Kopien unwiederbringlich verloren sind.

Was als archivierte Werk, bzw. Werkausprägung zu gelten hat, definiert aus Sicht des Archivs die Bestelleinheit. In der Nomenklatur der oft fälschlich als Standard bezeichneten – bei einem Standard muss man entscheiden können, ob etwas zum Standard konform ist – Empfehlung zum Offenen Archiv Informationssystem (OAIS) wird eine Bestelleinheit als Archival Information Package (AIP) bezeichnet. Ein eingereichtes Submission Information Package (SIP) wird mittels Normalisierung in ein AIP verwandelt und als solches im Archiv aufbewahrt. Bei Bestellung wird es mittels Denormalisierung in ein Dissemination Information Package (DIP) verwandelt und ausgeliefert. Damit ein solches „Paket“ in einem verteilten Repositorium bestellt werden kann, muss es mit einer weltweit eindeutigen ID identifiziert werden können. Eine solche ID muss offensichtlich von der konkreten Institution, wo das Werk archiviert ist, unabhängig sein.

Für verschiedene Objekte gibt es heute schon eine grosse Vielfalt von eindeutig kennzeichnenden IDs. Zum Beispiel die ISBN-Nummer für Bücher, die ISSN, die EAN, und so weiter. Die Proliferation solcher Nummern ist begleitet von einer Proliferation von Registrierungsinstitutionen, die versuchen, sich ein weltweites Monopol und damit viel Macht zu sichern. Traditionell sind diese IDs oft kompliziert strukturiert und lang, weil sie gleichzeitig mit der Identifikation eine Art Kurzbeschreibung des Werks und seines Standorts enthalten.

Moderne Datenbank-Theorie hat dagegen gezeigt, dass es von Vorteil ist, die Aufgabe der Beschreibung von der Aufgabe der Identifikation zu trennen und nur IDs zu verwenden, die keinerlei weitere Bedeutung tragen. Das hat etwa den Vorteil, dass die ID sich nicht ändert, wenn man einen Fehler in der Beschreibung entdeckt. So war früher in der „alten“ Schweizer Sozialversicherungsnummer das Geschlecht, der Geburtstag, die Anfangsbuchstaben des Namens und die Nationalität der Person kodiert. Das führte zu einer Änderung der Sozialversicherungsnummer, wenn sich Geschlecht, Geburtstag, Name oder Nationalität änderten und durchlöcherte den Zweck der eindeutigen Identifikation. Deshalb kodiert die „neue“ Sozialversicherungsnummer keine Beschreibungselemente mehr.

Alle heute verbreiteten Registrierungsmechanismen von Dokumenten eignen sich nicht für CultLib, weil sie einerseits mit hohen Transaktionskosten

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.

verbunden sind und andererseits nicht einheitlich für alle zu archivierenden Dokumente einsetzbar sind. Wir schlagen hier deshalb einen neuen dezentralen Ansatz für die Identifikation von Archivgut vor.

2 Werke und ihre Ausprägung

Um dem Archivgut eine eindeutige ID zuzuordnen zu können, muss zuerst Klarheit herrschen, welche Objekte eigentlich zu identifizieren sind. In einer Bibliothek ist die Bestelleinheit ein Buch und nicht zwei Seiten aus einem Buch oder eine Sammlung von Büchern. Wenn die Bibliothek zwei Exemplare desselben Buchs besitzt, ist es im Allgemeinen nur als ein Buch im Katalog verzeichnet und der Besteller erhält einfach ein Exemplar, das gerade nicht ausgeliehen ist. In eigentlichen Archiven ist die Auslieferungseinheit ein Dossier. Die Bestellung einer einzelnen Unterlage aus dem Dossier ist ebenso unmöglich, wie die Bestellung einer einzelnen Seite aus einem Buch bei einer Bibliothek.

2.1 Werkausprägung und Werkexemplar

Im Urheberrecht wird eine Unterscheidung zwischen dem immateriellen Werk und einem materiellen Werkexemplar gemacht. Dabei können sich Werkexemplare desselben Werks so ähnlich sein, wie zwei Exemplare derselben Auflage desselben Buchs oder so verschieden wie eine Luxusausgabe und eine Reclamausgabe eines Dramas von Schiller.

Auf den ersten Blick könnte man meinen, dass es keine digitalen Werkexemplare geben kann, denn es besteht kein Unterschied mehr zwischen Kopie und Original und jede Datei ist „immateriell“. Dennoch gibt es sehr wohl unterschiedliche digitale *Ausprägungen* (auf Englisch benutzen wir den Ausdruck „representation“) eines Werks. Diese unterscheiden sich so, wie die Luxusausgabe von der Reclamausgabe. Beispiele für eine Ausprägung sind: die digitale Verkleinerung einer Bilddatei, die aufgemotzte farbkorrigierte Version einer alten Fotografie, der bereinigte „Clean Master“ einer alten, kratzigen Tonaufnahme, die Restauration eines zerfallenden Films, eine Neuausgabe eines Textdokuments mit abweichendem Zeilen- und Seitenumbruch in einer anderen Schrifttype gesetzt. Aber es gibt auch Ausprägungen minderer Qualität, wie sie vor allem mittels Denormalisierung als Dissemination Information Package (DIP) aus dem Archivgut entstehen, wie Verkleinerungen, verlustbehaftete Kompressionen in Streamingformaten etc.

Im Projekt CultLib gehen wir davon aus, dass eine Werkausprägung immer eine einzelne Datei ist. Vorläufig sind keine „Dossiers“ im Repositorium von CultLib geplant. Um „Dossiers“ oder „Mail mit Anhängen“ sinnvoll in einem Repositorium lagern zu können, müsste erst eine standardisierte ZIP-Datei-Struktur für Dossiers oder Mails definiert werden. Diese könnten dann mit einer ID in den eingebetteten Metadaten identifiziert werden.

Ein Werk ist also eine Abstraktion der „konkreten“ Ausprägungen in Dateien, wenn es erlaubt ist, von Bit-Strömen als etwas Konkretes zu reden.

2.2 Ausprägung und Abwandlung

Letztlich ist es in Grenzfällen eine subjektive Entscheidung, ob ein Werk wie eine Übersetzung eines Romans ein von einem anderen Werk abgeleitetes Werk mit eigenem Werkcharakter – im Schweizer Urheberrecht heisst das „Werk zweiter Hand“, in der Schweizer Version der Creative Commons Lizenz „Abwandlung“ – oder „nur“ eine Ausprägung desselben Werks ist.

Masstab für diese Abgrenzung ist wohl das Ausmass an kreativ Neuem, welches zum abgewandelten Werk hinzutritt. Keine Rolle spielt dagegen die Mühe, die es gemacht hat, eine Ausprägung herzustellen. Digitalisate analoger Originale in hoher Qualität sind oft mühsam herzustellen. Sie begründen aber grundsätzlich kein neues Werk.

2.3 Sind Aufführungen bloss Ausprägungen?

Im Urheberrecht ist es üblich, eine musikalische Darbietung als Werkexemplar, als blosse Ausprägung des Werks des Komponisten zu behandeln. Analog sind Regisseur und Darbieter eines Theaterstücks oder eines Films keine Urheber im juristischen Sinn. Dieses Privileg bleibt Drehbuchautoren und Komponisten vorbehalten. Diese Anomalie kann man leicht verstehen, wenn man in Betracht zieht, dass das Urheberrecht europäischer Prägung aus dem Streit von Theaterautoren und Schauspielern hervorgegangen ist. Auch die Verwertungsgesellschaften entstanden aus dem Streit zwischen Textern und Komponisten einerseits und aufführenden Musikern andererseits, der von einem französischen Gericht zu Ungunsten der ehrlich arbeitenden Ausführenden entschieden wurde.

Im Zusammenhang mit der Frage der Langzeiterhaltung präsentiert sich die Angelegenheit anders. Das Produkt eines Drehbuchautors ist ein Drehbuch. Das Produkt des Regisseurs und der Filmcrew ist ein Film. Diesen mag man gerne als Abwandlung, als abgeleitetes Werk des Drehbuchs bezeichnen. Ihm aber völlig den eigenständigen Werkcharakter abzusprechen geht wohl nur in den Gehirnen von Kulturfunktionären auf. Im Rahmen von CultLib bezeichnen wir jedenfalls einen Song als Tondatei als ein eigenständiges Werk, das auf dem Notentat des Komponisten als Vorbild beruhen kann. Auch in der kulturellen Praxis sind ja heute Darbietung und Komposition von Songs wieder näher zusammengerückt.

Wenn man die juristische Definition einer Abwandlung weit genug fasst, ist wohl fast jedes Werk ein Werk zweiter Hand.

2.4 Sind Sammlungen auch Werke?

In diesem Bericht wird versucht, einige für die Langzeiterhaltung wichtige Begriffsunterscheidungen anhand von Beispielen plausibel darzustellen. Diese Begriffsklärung soll ein Leitfaden für die Arbeit am digitalen Repositorium sein. Wir nehmen uns nicht heraus, für die ganze Welt festzulegen, was ein Werk ist und wie ein zu normalisierendes Submission Information Package (SIP) auszusehen habe. Solche normative Empfehlungen suche man besser im OAIS-Dokument.

Ob Sammlungen auch Werke sind, soll daher hier nicht normativ entschieden werden. Es scheint aber einfacher und pragmatischer, den Status von Sammlungen nicht zu regulieren. In CultLib ist die Bestelleinheit immer die Dateiausprägung eines Werks. Wer eine Sammlung bestellen will, muss mehrere Werke bestellen. Der Sammlung selber sprechen wir - eventuell unge rechterweise - den Werkcharakter ab und lassen auch offen, ob es Sammlungen von Sammlungen geben darf. Ein Inhaltsverzeichnis und Konzept einer Sammlung kann hingegen durchaus die Sammlung vertreten und Werkcharakter haben.

3 Identifikation von Werk und Werkausprägung

Es werden also IDs für Werke und ihre Ausprägungen benötigt, wenn man ein digitales Archiv aufbauen will. Als weltweit eindeutige ID für das abstrakte Werk bietet sich der sogenannte Universally Unique Identifier (UUID), manchmal - nur wenig bescheidener - auch Globally Unique Identifier (GUID³) an. Diese haben den Vorteil, dass sie von jedermann ohne eine zentrale Registrierungsinstanz erzeugt werden können und garantiert eindeutig sind. Bei der UUID handelt es sich um einen freien, offenen Standard.

Zur Erzeugung einer solchen UUID genügt es auf vielen Betriebssystemen, einfach den Befehl *uuidgen* auszuführen. Es gibt auch viele Online-Generatoren wie etwa <http://www.guidgen.com/>. Auf der Windows-Plattform scheint der entsprechende Generator verschwunden zu sein. Wer eine JAVA-Plattform zur Verfügung hat, kann das Open Source Werkzeug *UuidGen* verwenden.

Eine solche UUID ist natürlich nicht im ganz strikten mathematischen Sinn eindeutig. Es ist nur äusserst unwahrscheinlich, dass eine Kollision stattfindet - auch wenn alle Menschen der Welt alle Sekunden eine neue ID generieren. Jedenfalls ist eine Doppelerzeugung derselben ID sehr viel unwahrscheinlicher als eine Kollision bei einer Registrierungsinstanz. Und auch bei einer solchen werden menschliche Fehler gemacht und müssen manuell korrigiert werden.

Als Identifikation der Ausprägung (representation) eines Werks, also einer konkreten Datei, bietet sich ein sogenannter Message Digest oder kryptographische Hashfunktion an. So nennt man eine Art Quersumme über alle Bytes in einer Datei, die für jede Datei einen anderen Wert hat. Da ein solcher Digest typischerweise kurz ist (MD5: 16 Bytes bzw. 32 Hexadezimalziffern; SHA-1: 20 Bytes bzw. 40 Hexadezimalziffern) können natürlich unmöglich alle denkbaren Dateien jeder Länge einen unterschiedlichen Digest erzeugen. Aber wieder gilt, dass die Wahrscheinlichkeit, dass zwei verschiedene Dateien zum selben Digest führen, äusserst klein ist. Darum wird ein solcher Digest auch als Fingerabdruck der Datei bezeichnet. Wir gehen davon aus, dass in Zukunft neben den heute verbreiteten MD5, SHA-1, SHA-256 weitere Digestmethoden populär werden können. Wir lassen alle freien, offenen Standards für einen Message Digest zu, und versehen ihn mit einem Präfix, der die Methode identifiziert. Zum Beispiel:

SHA-1:d253e3bd69ce1e7ce674345fd5faa1a3c2e2030ef,

MD5:57edf4a22be3c955ac49da2e2107b67a.

Um eine Werkausprägung zu identifizieren nehmen wir also einen Digest über die Primärdaten. Denn wir wollen den Digest selber in den eingebetteten Metadaten speichern und ausserdem von derselben Ausprägung eines Werkes reden, auch wenn die Metadaten verändert wurden. Während der Archivierung dürfen Primärdaten nie verändert werden. Metadaten hingegen schon. Bei Containerformaten wie WebM oder OOXML, die selber wieder selbständige Dateien mit eigenen Metadaten enthalten, kann man den Digest über die Digests der Primärdaten der einzelnen Inhaltsdateien berechnen. Wie der

³ Der technische Unterschied zwischen UUID und GUID, nämlich ob die am meisten oder am wenigsten signifikanten Bits zuerst kommen, ist mindestens bei den heute am weitesten verbreiteten Zufalls-IDs irrelevant.

Message Digest konkret zu berechnen ist, wird für jedes einzelne Format explizit festgelegt. So kann der Digest jederzeit anhand der Primärdaten neu berechnet werden. Das kann zur Überprüfung der Authentizität der Primärdaten verwendet werden.

Damit ein Werk wirklich über eine UUID und jede seiner Ausprägungen über einen Digest identifiziert wird, ist es notwendig, dass bei der Herstellung von Ausprägungen eingebettete Metadaten auf die neue Ausprägung übergehen. Bei eingebetteten Metadaten ist dies heute schon in vielen Fällen automatisch der Fall. Es würde die eindeutige Identifikation eines Werkes erschweren, wenn mehrere IDs für dasselbe Werk verwendet werden. Es wird sich aber, besonders auch bei Digitalisaten, sicher nie ganz vermeiden lassen. Solange nicht dieselbe ID für verschiedene Werke verwendet wird, lässt sich damit leben.

4 Generelle Pflicht zur Identifikation?

Wenn es für Urheber generelle Pflicht wäre, ein neu erstelltes digitales Werk mit einer ID auszuzeichnen, würden viele Ärgernisse des heutigen Urheberrechts beseitigt. Es wäre nicht mehr, wie heute, automatisch jede Datei urheberrechtlich geschützt. Kein Richter müsste über die „Schöpfungshöhe“ eines Werks philosophieren, welche erreicht sein muss, damit urheberrechtlicher Schutz greift. Stattdessen stünde es im Belieben der Urheber, diejenigen Objekte auszuzeichnen, die sie als Werke unter urheberrechtlichen Schutz gestellt haben wollen. Dazu müssten sie sich mit keiner Registrierungsinstanz herumärgern und keine Gebühren bezahlen. Ein Eintrag einer selbst generierten ID in den eingebetteten Metadaten genügt. Und ein solcher ist auf allen modernen Plattformen sehr einfach zu bewerkstelligen.

Eine solche ID kann zu irgendeinem Zeitpunkt – am besten vor der Veröffentlichung – hinzugefügt werden. Als Veröffentlichung ist diejenige Aktivität der Urheber anzusehen, die einer anonymen Menge von Personen das Kopieren ermöglicht. Die juristischen Unterschiede zwischen Vortragen, Aufführen, Publizieren, Veröffentlichenden, Zugänglichmachen etc. entsprechen im digitalen Zeitalter ohnehin keinen realen Gegebenheiten mehr und können unter dem einzigen Begriff Publikation zusammengefasst werden.

Wenn man den „Lebenszyklus“ eines Werkes betrachtet, ist die UUID die übergeordnete Identifikation eines Werks, während der Message Digest dessen Version eindeutig bezeichnet. Wir erwarten, dass sich die Versionen nach Publikation eines Werks wenig und langsamer als vor der Publikation verändern. Wenn ein Werk im Archiv landet, sind nur noch wenige, kleine Änderungen zu erwarten.

Zusätzlich zur ID sollten die Urheber die Möglichkeit haben, das Publikationsdatum in den eingebetteten Metadaten festzuhalten. Ein fehlendes Publikationsdatum kann dann darauf hinweisen, dass etwas noch nicht publik ist und als Teil der Privatsphäre geschützt ist, während eine fehlende ID bei einem vorhandenen Publikationsdatum darauf hinweist, dass kein urheberrechtlicher Schutz beansprucht wird, weil das Werk im Auge des Urhebers keine genügende Schöpfungshöhe erreicht.

Wenn nun auch noch die Lizenzbestimmungen in den eingebetteten Metadaten festgehalten sind, hat jeder Nutzer die Möglichkeit, den legalen Status und somit die Nutzbarkeit einer Werkausprägung festzustellen. Die heute grassierende Urheberrechts-Fallenstellerei mit unklarem urheberrechtlichem Status wäre verunmöglicht.

5 URL-Struktur für die Anforderung von Dokumenten

Im Rahmen der Standardisierung des World Wide Web wurden immer wieder Anläufe unternommen, um eine von der Lokalität unabhängige Identifikation zu definieren. So ging etwa aus der URL, dem universellen Ressourcen-Lokator die URI, die universelle Ressourcen-Identifikation hervor. Ganz sauber trennte sich diese aber nicht von der URL, weil sie nicht bei der Lokalisierung hilft und weil die Registrierung neuer URIs aufwendiger ist als die Registrierung einer neuen Domäne.

Wer konkret ein Werk von einem Repositoryum - etwa mit Hilfe eines Browsers - anfordern will, muss ja zwangsläufig eine konkrete Adresse mit einem Domännennamen verwenden. Hier legen wir fest, wie die oben definierten IDs in URLs einzubinden sind:

Mit

<protokoll>://<domäne>[:<port>]/<repositoryum>/cultlib/<uuid>

fordert man eine - die qualitativ höchststehende verfügbare - Ausprägung des durch die UUID identifizierten Werks an. Dabei steht

- <protokoll> für „http“ oder „https“ oder zukünftige Protokollbezeichnungen,
- <domäne> für einen Namen in einer registrierte Domäne wie etwa „www.enterag.ch“,
- <port> für die Portnummer; nur nötig, wenn eine vom Standardport für das Protokoll abweichende Nummer verwendet wird,
- <repositoryum> für einen Pfad zur Startseite des Repositoryums,
- <uuid> für die ID des Werks.

Ein Beispiel für eine solche Werk-URL ist

<http://www.cultlib.ch/repositoryum/cultlib/88e836cf-c830-44e9-9e60-0a08e0aba500>.

Analog fordert man eine konkrete Ausprägung mit der URL

<protokoll>://<domäne>[:<port>]/<repositoryum>/cultlib/<uuid>/<digest>
an.

Ein Beispiel für eine URL einer Werkausprägung ist

<http://www.cultlib.ch/repositoryum/cultlib/88e836cf-c830-44e9-9e60-0a08e0aba500/MD5:57edf4a22be3c955ac49da2e2107b67a>.

Anstelle des Werks kann man auch nur die Metadaten eines Werks oder einer Werkausprägung anfordern.

<http://www.cultlib.ch/repositoryum/cultlib/metadata/88e836cf-c830-44e9-9e60-0a08e0aba500>

oder

<http://www.cultlib.ch/repositoryum/cultlib/metadata/88e836cf-c830-44e9-9e60-0a08e0aba500/MD5:57edf4a22be3c955ac49da2e2107b67a>.

Für autorisierte Mitarbeiter des Archivs dient dieselbe URL zum Ändern der Metadaten eines Werks, indem statt dem HTTP-Befehl GET der Befehle POST verwendet wird und die abgeänderten Metadaten im XML-Format als Inhalt der HTTP-Abfrage übermittelt werden. Ausführlicher werden die eingebetteten Metadaten im nächsten Bericht zur Langzeiterhaltung dargestellt.

Ganz im Sinne des Konzepts der Webservices, die auf dem Konzept des Representational State Transfer (REST) beruhen, sind solche URLs *permanent*.

Das heisst, dass sie immer dieselbe Ressource – allenfalls mit korrigierten Metadaten – liefern, sofern der Server noch in Betrieb ist und die Ressource überhaupt noch liefern kann.

5.1 URLs mit Abfragen

Es gibt Fälle, wo man den Message Digest einer Ausprägung nicht kennt, aber umschreiben kann, was man abzuholen wünscht. In diesem Fall benutzt man Abfragen (Queries), die in der URL hinter einem Fragezeichen aufgeführt werden. Im Gegensatz zu URLs ohne Fragezeichen sind Abfrage-URLs nicht im selben Sinne „permanent“. Dieselbe URL kann zu verschiedenen Zeitpunkten unterschiedliche Resultate haben.

Die Syntax solcher Queries wird hier nicht reguliert und kann vom einen zum anderen Repository unterschiedlich sein. Ein paar Beispiele illustrieren, wie man Queries in URLs handhaben kann:

<http://www.cultlib.ch/repository/cultlib/88e836cf-c830-44e9-9e60-0a08e0aba500?scale=0.5>

für Anforderung eines um den Faktor 0.5 verkleinerten Version eines Bilds.

<http://www.cultlib.ch/repository/cultlib/88e836cf-c830-44e9-9e60-0a08e0aba500?height=456&width=867>

für die Anforderung eines skalierten Bilds mit 456 Pixels Höhe und 867 Pixels Breite.

<http://www.cultlib.ch/repository/cultlib/88e836cf-c830-44e9-9e60-0a08e0aba500?format=jpeg&compression=50>

für die Anforderung einer denormalisierten Version des Bilds als JPEG-Datei.

In all diesen Fällen erhält man eine Ausprägung, ein Dissemination Information Package (DIP), welches eine Kopie der eingebetteten Metadaten der qualitativ besten Ausprägung des Bilds als eingebettete Metadaten enthält. Man kann auch analog die Ableitung von einer spezifischen Ausprägung anfordern.

Die genaue Abfragesyntax ist von jedem Repository für jedes unterstützte Format festzulegen.

5.2 URLs und Metadaten

Die standardisierten Metadaten können für die Suche nach einem Werk benutzt werden.

Auch diese Syntax soll hier nicht reguliert, sondern nur exemplarisch angedeutet werden:

<http://www.cultlib.ch/repository/cultlib?creator=Hartwig&title=Langzeiterhaltung>

würde eine Liste aller Metadatenätze von Dokumenten im Archiv anfordern, deren Metadatenfeld „creator“ die Zeichenkette „Hartwig“ und deren Titel in irgendeiner Sprache „Langzeiterhaltung“ enthält.

5.3 URLs und andere Listen

Die Suche nach Archivgut kann auch vielfältig nach Sammlungen oder anderen Aspekten der Tektonik des Archivs eingeschränkt werden. Solche Abfragen sind archivspezifisch. Andere Archive werden dasselbe Werk eventuell nicht in einer gleichnamigen Sammlung gespeichert haben.

Ein Beispiel:

<http://www.cultlib.ch/repository/cultlib?collection=Leuzinger&type=video>

würde eine Liste der Metadatensätze aller Videodokumente in der Sammlung Leuzinger anfordern.