Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

ANTJE DITTMANN[1], TIM HASLER[2], ELIAS OLTMANNS[3]

Archivierungsstrategie für audiovisuelles Material

[1]Stiftung Deutsche Kinemathek
[2] 0000-0001-9164-3500, Zuse-Institut Berlin
[3] 0000-0003-1453-7063, Zuse-Institut Berlin

ZIB Report 22-03 (Februar 2022)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30 84185-0
Telefax: +49 30 84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Archivierungsstrategie für audiovisuelles Material

Antje Dittmann (Deutsche Kinemathek)
Tim Hasler (Zuse-Institut Berlin)
Elias Oltmanns (Zuse-Institut Berlin)

Inhalt

Abstract

Im Rahmen ihrer Strategie zur Langzeitarchivierung forschte die Deutsche Kinemathek in einer Kooperation mit dem Zuse-Institut Berlin (ZIB) an der digitalen Langzeitarchivierung von AV-Materialien. Ausgangspunkt des Projektes waren die enormen Dateigrößen und die heterogenen Dateiformate, die einem Werk und einer Fassung zugeordnet werden müssen. Die Verwendung von persistenten Identifikatoren stellt den Lösungsansatz dar. Das beschriebene Projekt lief von 01/2019–12/2021.

Ausgangslage

Filme[1] existieren seit ihren frühesten Anfängen in unterschiedlichen Schnitt- und Materialfassungen. Diese verschiedenen Fassungen werden in der Distribution weitergeführt und liegen auch so in den Archiven vor. Verschiedene Institutionen halten zudem oft verschiedene Materialien zu demselben Filmwerk. Bei der Digitalisierung von Filmmaterial im Rahmen von Restaurierungsvorhaben erweitern sich diese analogen Objekte um Datenobjekte wie Bildsequenzen, Audiodateien,

1 „Mit fotografischen oder elektronischen Mitteln erzeugte Folge von Einzelbildern, die – auf eine Leinwand projiziert oder auf einem Bildschirm sichtbar gemacht – den Eindruck von Bewegung hervorruft.“ http://d-nb.info/gnd/4017102-4 [23.06.2021].

Untertiteldateien und Distributionsformate, die oft zeitversetzt anfallen. Bei jeder Digitalisierung entsteht eine weitere Fassung eines audiovisuellen Werkes, die die bereits vorliegenden ergänzt.

Es werden sowohl Mastermaterialien (Bildsequenzen und Töne in ihrer unbearbeiteten Form) erhalten, als auch Distributionsformate wie DCPs[2] bestehend aus Bildern, Tönen und Untertiteln sowie Bilddateien archiviert. Eine gebräuchliche Erfassung der Datenobjekte in den Filmarchiven ist die Ablage und Verzeichnung in materialspezifischen Datenpaketen. Dabei muss sowohl der Bezug der Datenobjekte untereinander, als auch der auf die analoge Vorlage und auf ein übergeordnetes Werk erhalten bleiben. Das wird in der Regel durch die institutionseigenen Erschließungssysteme sichergestellt. In der Praxis stellen die Datenobjekte die Handlungseinheit bei Kooperationen zwischen filmhaltenden Institutionen dar.

Zusätzlich zu den AV-Materialien bewahren und erschließen viele Filminstitutionen die sogenannten filmbegleitenden Objekte. Hierbei handelt es sich um Produktionsunterlagen, Fotos, Plakate, Kostüme, Technik und vieles mehr. Diese Objekte dienen als Forschungsinfrastruktur und sind damit Grundlage der filmwissenschaftlichen Forschung. Sie ermöglichen interdisziplinäre Forschung, bei denen die gesellschaftlichen Hintergründe der Filmentstehung, technische Praktiken und soziologische Zusammenhänge untersucht werden. Diese heterogenen Sammlungsbestände werden zunehmend digitalisiert und nach objekttypischen Erschließungsstandards erfasst. Gemeinsam ist den Objekten die Notwendigkeit, sie durch eindeutige Verweise als Bestandteil einer bestimmten Fassung oder eines darzustellenden Werkes idealerweise über die Grenzen der eigenen Institution hinaus kenntlich zu machen, um die wissenschaftliche Nutzung auch im globalen Zusammenhang zu ermöglichen.

Die Verzeichnung filmografischer[3] Daten an der Deutschen Kinemathek wie auch an einigen internationalen Filmarchiven folgt dem europäischen Standard CEN TC 372.[4] Dieser wurde ab dem Jahr 2005 entwickelt, um den Zugang zum europäischen Filmerbe zu erleichtern.[5] Der Standard teilt sich in ein Minimum-Set (EN 15744)[6] zur Identifizierung audiovisueller Werke und in ein Maximum-Set (EN 15907)[7], bestehend aus Elementen und Strukturen zur Erhöhung der Interoperabilität von Metadaten. Das Datenmodell von EN 15907 basiert auf FRBR[8] und umfasst als hierarchische Entitäten Werk, Variante, Manifestation und Item. Das vierstufige Modell kann optional, wie an der Deutschen Kinemathek praktiziert, auch dreistufig in Werk, Manifestation und Item umgesetzt werden. Dabei fasst die Entität Manifestation den Produktions- und Publikations-Event zusammen und verzichtet auf die separate Darstellung der Variante. Im vorliegenden Bericht wird der allgemeinere Begriff Fassung gebraucht, der in diesem Kontext beide Entitäten (Variante und Manifestation) meint.

2 Digital Cinema Package (DCP) ist das vorherrschende Containerformat zur Projektion von digitalen Kinofilmen.
3 Filmografisch steht für beschreibend, deskriptiv.
4 https://standards.iteh.ai/catalog/tc/cen/6c8e269b-f49c-4cdd-b636-11693960c3c6/cen-tc-372 [05.10.2021].
5 https://www.ace-film.eu/?page_id=152 [05.10.2021].
6 http://filmstandards.org/fsc/index.php?title=EN_15744 [05.10.2021].
7 http://filmstandards.org/fsc/index.php?title=EN_15907 [05.10.2021].
8 FRBR (Functional Requirements for Bibliographic Records) steht für ein durch die IFLA (International Federation of Library Associations and Institutions) entwickeltes Datenmodell für bibliothekarische Metadaten.

Nach der Einführung 2010 in der europäischen Filmarchiv-Community wurde deutlich, dass fehlende Interpretationshilfen zu unterschiedlichen Auslegungen des Standards führten. Eine Hilfestellung stellte das 2016 erschienene *The FIAF Moving Image Cataloguing Manua*l dar, welches eine Harmonisierung der Erfassungsstandards FRBR, RDA[9] und EN 15907 zum Ziel hatte und auf die speziellen Bedürfnisse der filmhaltenden Institutionen eingeht.[10] Obwohl das Manual versucht, den Rahmen des CEN TC 372 regelkonform zu füllen, blieben in Bezug auf die Erfassungsrichtlinien viele Optionen offen. In der Praxis behindern diese unterschiedlichen Auslegungen das Ziel der Interoperabilität. Die größte Divergenz gibt es bei der Zuordnung zu Variante oder Manifestation, die schon durch die optionale Drei- oder Vierstufigkeit des Datenmodells angelegt ist. Die in mehreren Filminstituten genutzte Datenbanksoftware Adlib der Firma Axiell, setzt das dreistufige Datenmodell um, und war lange Zeit die einzige kommerzielle Software, die den Standard abbildete.

Eine weitere Ursache der disparaten Umsetzung des filmografischen Standards ist das Fehlen eines standardisierten XML-Schemas der Metadatensets. In Ermangelung eines spezifizierten Serialisierungsformats für CEN TC 372, in dem die Metadaten gespeichert und ausgetauscht werden können, entwickelten Institutionen eigene XML-Umsetzungen für EN 15744 oder nutzen Elemente des EN 15907 in proprietären Austauschschemata. So stellt das Portal *finna.fi* zu den Datenbankeinträgen der Finnischen Nationalen Filmografie einen XML-Auszug nach EN 15907 zur Verfügung, dessen Schema im EU-Projekt *Forward*[11] entwickelt wurde.[12] Auch das Deutsche Filminstitut Filmmuseum/Filmportal entwickelte ein eigenes Schema für EN 15907[13] und ein *German Archival Data Exchange Schema*.[14] Durchsetzen konnte sich bisher keines dieser Schemata.

Mit der zunehmenden Strukturierung, Standardisierung und Publikation der Metadaten wird der Bedarf an Normdaten und an persistenten Identifikatoren für Filmwerke und für Fassungen dringlicher.[15] Während für Personen, Körperschaften, Geografika und zunehmend auch Schlagwörter Normdaten und Identifikatoren zum alltäglichen Verzeichnungsprozess von AV-Materialien gehören, stellen Identifikatoren für Werke und Fassungen, die die Bedarfe der filmhaltenden Institutionen treffen, ein großes Desiderat dar. Der menschliche Nutzer kommt bei Suchanfragen auch mit unpräzisen Anfragen zu verwertbaren Ergebnissen, doch Maschinen benötigen präzise Angaben für ein verlässliches Ergebnis. So werden die Regeln zur Darstellung von Titeln audiovisueller Werke und ihrer Fassungen in Bibliotheken häufig ignoriert, was eine automatisierte Identifikation und Zuordnung erschwert bis unmöglich macht.[16] Persistente, globale Identifikatoren sind daher nicht nur für die eindeutige Identifikation eines Werkes nötig, sondern sie ermöglichen ein Netzwerk

9 RDA (Resource Description and Access) ist ein Regelwerk, das auf Basis des FRBR-Konzeptes entwickelt wurde.
10 https://www.fiafnet.org/pages/E-Resources/Cataloguing-Manual.html, *1-2* [19.10.2021].
11 https://cordis.europa.eu/project/id/325135, *Framework for a EU-wide Audiovisual Orphan Works Registry* [19.10.2021].
12 https://elonet.finna.fi/Record/kavi.elonet_elokuva_120963 [18.10.2021].
13 https://filmstandards.org/schemas/EN15907-d1/ [18.10.2021].
14 https://filmstandards.org/schemas/de-dif/fw-exch-1.0/ [18.10.2021].
15 Stoppe, *Streaming für Forschende – Desiderata aus Sicht des Fachinformationsdienstes für Kommunikations-, Medien- und Filmwissenschaft*, *2017*, https://doi.org/10.1515/bfp-2020-2041 [09.12.2021].
16 Kroon, Drewery, Leigh und McConnachie, *Content Identification for Audiovisual Archives*, *2015,* 21, International Association of Sound and Audiovisual Archives (IASA) Journal, (45), 20–30, https://doi.org/10.35320/ij.v0i45.80.

zwischen AV-Beständen in ihren diversen archivarischen Ausprägungen über einzelne Institutionen hinweg.

Forschungskooperation Deutsche Kinemathek & Zuse-Institut Berlin

Aufgrund des hohen technologischen Aufwandes wird an der Deutschen Kinemathek davon ausgegangen, dass eine qualifizierte Langzeitarchivierung nicht durch die Institution selbst umgesetzt werden wird. Die Archivierung der Digitalisate ist immer als parallele Aufgabe zur konservatorischen Pflege der analogen Materialien zu sehen und verursacht in der Regel sogar höhere Kosten[17].

Das bedeutet für viele Filmarchive, speicherintensive Formate wie Masterdateien[18] werden mittelfristig an etablierte, OAIS-konforme[19] Langzeitarchive abgegeben und nur Distributionsformate im Haus vorgehalten.[20] Die mitunter zeitlich verteilt eintreffenden Datenobjekte müssen daher zukünftig kontinuierlich in die Langzeitarchivierung überführt werden können, während die Zugehörigkeit zu Werk und Fassung und zur jeweiligen Repräsentation im Erschließungssystem bewahrt wird.

In der Ausgangssituation zu Projektbeginn wurden die Datenobjekte der Deutschen Kinemathek in einzelnen Verzeichnissen auf der eigenen Tape Library archiviert und konnten durch eine interne Datenbankabfrage identifiziert und zurückgespeichert werden. Die Verzeichnung der Metadaten in der Sammlungsdatenbank Adlib erfolgt wie beschrieben dreistufig. Aufgrund der großen Datenmengen und der archivinternen Abläufe ist es wenig praktikabel, alle Datenobjekte einer Fassung – oder sogar die der obersten Entität Werk – gleichzeitig mit allen vorhandenen Metadaten – für die digitale Langzeitarchivierung zu einem großen physischen Submission Information Package (SIP) zusammenzufassen. Obwohl es technisch möglich wäre, auch Datenmengen im Petabyte Bereich zu organisieren und zu transportieren, ist dies eher die Domäne von Rechenzentren. Bei Abgabe von einzelnen Datenobjekten wie beispielsweise Mastermaterialien an ein Langzeitarchiv oder bei einem Austausch von Daten im Rahmen von Kooperationen zwischen filmhaltenden Institutionen, sind die Datenobjekte der Deutschen Kinemathek nicht ohne Weiteres einem Werk oder einer Fassung zuzuordnen. Beigelegte Exporte der einordnenden Metadaten als proprietäre XML sind nicht automatisiert verifizierbar. Im Fall eines Accessprozesses[21] nach einer Abgabe an ein Langzeitarchiv und der Rückgabe an die Kinemathek, lassen sich diese proprietären Metadaten-Dateien nach einem gewissen Zeitraum nicht mehr vollständig konfliktfrei in die eigene, sehr dynamische Institutionsdatenbank einlesen. Als Ergebnis der Analyse von Verzeichnungs- und Datenhaltungsprozessen wurde durch das ZIB die Einführung von persistenten Identifikatoren bei der Archivierung von AV-Daten empfohlen.

---

17 Auch wenn ein direkter Kostenvergleich schwierig ist, wird durch die zunehmend bereits digital erzeugten Aufnahmen der Anteil der digitalen Objekte mittelfristig den größten Teil der Bestände ausmachen.
18 Gemeint sind hier unkomprimierte Formate wie DPX (Digital Picture Exchange) oder TIFF 16 bit (Tagged Image File Format).
19 http://www.oais.info/ [09.12.2021], siehe auch TRAC (Trustworthy Repositories Audit & Certification) https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf [05.10.2021].
20 Ein wesentliches Desiderat im Preservation Management stellt weiterhin die Etablierung eines stabilen, offenen Archivformates und Archivcodecs für AV-Daten dar.
21 Zu Access-Prozess siehe OAIS-Modell, http://www.oais.info/ [10.12.22021].

PID

Ein persistenter Identifikator (PID) bezeichnet ein Objekt eindeutig und dauerhaft. Hierfür sind prinzipiell auch institutionelle Identifikatoren wie URIs[22] denkbar. Dabei obliegt die Auflösung der Identifikatoren durch Metadaten dann der jeweiligen Institution mit ihrer Infrastruktur. In jedem Fall muss ein institutioneller Identifikator als proprietäres System angesehen werden, das in der Regel anderen Institutionen nicht zur Verfügung steht. Die Vergangenheit hat gezeigt, dass URLs[23] als Teilmenge der URIs wenig dafür geeignet sind, da sich häufig Webseitenstrukturen oder auch Domainnamen über die Zeit ändern. Es gibt daher eine Entwicklung zu institutionsübergreifenden, globalen, persistenten Identifikatoren, die nicht einer Institution allein "gehören", sondern konsortial genutzt werden. Diese Identifikatoren lassen sich öffentlich auflösen und haben globale Strategien zur Persistenz entwickelt.

In den Wissenschaften hat sich der Digital Object Identifier (DOI) etabliert, mit dem zunächst Publikationen von Verlagen eindeutig referenziert wurden und der zunehmend auch für Datensätze selbst Verwendung findet. Zur Erzeugung eines DOI müssen bestimmte Metadaten im DOI-System hinterlegt werden, die die referenzierte Ressource grundlegend beschreiben. Weiterhin ist die Angabe einer sogenannten Landing Page verpflichtend, auf der eine detaillierte Beschreibung der Ressource erfolgt. Diese Landing Page hat den Vorteil, auch noch Informationen liefern zu können, sollte das Objekt nicht mehr zugänglich sein. Über die Metadaten und die Landing Page wird die Persistenz des DOI über die Lebenszeit des Objekts hinaus sichergestellt. Sie kann auch eine Überprüfung der Legitimation des Zugriffs ermöglichen, etwa bei sozialwissenschaftlichen oder medizinischen Daten. Mit der Verzeichnung der Basismetadaten entsteht eine öffentlich zugängliche Datenbasis, die die Sichtbarkeit der referenzierten Objekte erhöht und eine Nachnutzung ermöglicht.

Der im Rahmen des Projektes entwickelte Lösungsansatz besteht darin, persistente Identifikatoren mit geeigneten Basismetadatensets einzusetzen, um Datenobjekte global identifizierbar, ihre Beziehung zueinander und zu einem Werk darstellbar zu machen. Damit sollen die Voraussetzungen für gut verstehbare Archivpakete und möglichst hohe Interoperabilität zwischen den Archiven geschaffen werden.

Im vorgeschlagenen PID-System werden drei Entitäten unterschieden:

1. Die abstrakte Entität *Werk (work)* wird aus EN 15744 übernommen.
2. *Fassung (version)* steht für das vorliegende audiovisuelle Material zur Darstellung einer Ausprägung eines Werkes, das archiviert werden soll
3. *Datenobjekt (dataObject)* steht für ein konkretes Archivpaket, wobei der Inhalt immer genau einer Fassung zuzuordnen ist.

Über wechselseitige Referenzen in den Basismetadatensets wird die Fassung mit den zugehörigen Datenobjekten verknüpft. Damit ist sichergestellt, dass die Fassung als vollständig versteh- und

22 URI (Uniform Resource Identifier) Identifikator zur Identifizierung einer abstrakten oder physischen Ressource.
23 URL (Uniform Resource Locator) Identifikator zur Identifizierung und Lokalisierung einer Ressource.

nutzbare Einheit aus den Datenobjekten rekonstruiert werden kann. Das dargestellte Werk wird mit seinen Metadaten aus EN 15744 ebenfalls im PID-System verzeichnet und von der Fassung referenziert. Somit entsteht aus den Metadaten, die zur Registrierung von PIDs gefordert werden, eine Datenbank, die mit jedem registrierten PID wächst. Alle Identifikatoren der Fassungen und Datenobjekte können einer Institution zugeordnet werden und ordnen sich selbst der abstrakten Werksebene zu. Diese Konstruktion erhöht die Transparenz über Archivgrenzen hinaus und bietet Vorteile im Hinblick auf zukünftige Dienste jenseits der Archivierung.

Zur Überführung in ein Langzeitarchiv enthält dann jedes Submission Information Package (SIP)[24] neben den audiovisuellen Daten (Datenobjekt) alle für dieses Objekt relevanten Metadaten. Das sind die PIDs sowohl für das Datenobjekt selbst, als auch die Fassung und das Werk samt aktuellem Stand der entsprechenden Basismetadaten. Weil die PIDs mit ihren registrierten Metadaten global, online auflösbar sind, besteht für die Community jederzeit die Möglichkeit, auch die Basismetadaten aller übrigen Datenobjekte der Fassung abzurufen.

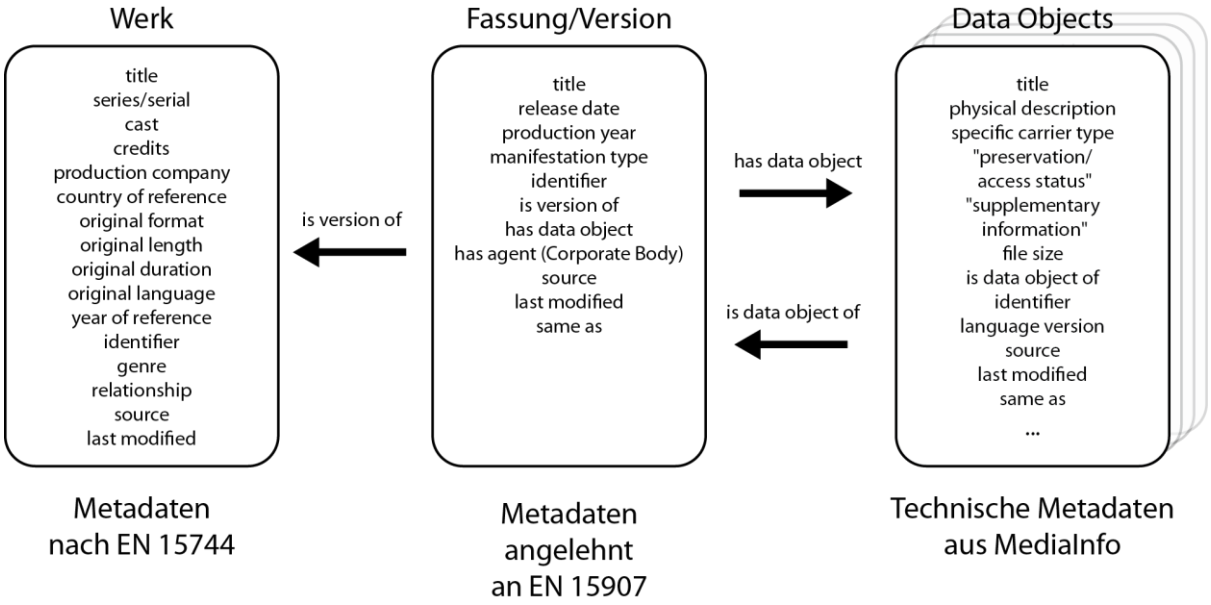

Abb. 1 Beziehungen der Objekte im PID-System

Ebenso besteht zukünftig die Möglichkeit, auch filmbegleitende Sammlungsmaterialien mit PIDs und somit einer definierten Relation zum verzeichneten Werk zu versehen. Gleiches gilt für die analogen Vorlagen. Die Referenzierung analoger Materialien durch PIDs ist bereits etablierte Praxis.[25]

Übersicht der PID-Systeme im AV Bereich

Zur Identifizierung von audiovisuellen Werken und Fassungen existieren zwei durch die Filmindustrie entwickelte Identifikatorensysteme. Die International Standard Audiovisual Number (ISAN) und die

24 Zu SIP (Submission Information Package) siehe OAIS-Modell, http://www.oais.info/ [10.12.2021].
25 Beispiele sind die Identifizierung von Diapositiven in der ETH Zürich (http://doi.org/10.3932/ethz-a-000074621), oder die International Geo Sample Number IGSN (https://www.igsn.org/about/), die als globales Referenzsystem initiiert wurden.

Entertainment Identifier Registry (EIDR). Die Nutzer kommen zum großen Teil aus dem Verwertungsbereich und beide Organisationen haben die Unterstützung der Filmindustrie zum Ziel.[26] Das EIDR- und ISAN-System sind seit 2019 interoperabel.[27] Bei ISAN handelt es sich um einen ISO-Standard[28], der 2007 um die Ebene der Versionen[29] erweitert wurde. Die Vergabe der Nummern erfolgt über nationale Registrierungsagenturen und der Preis für die Registrierung eines Identifiers liegt bei € 16,- was die Nutzung bis hinunter auf Objektebene unpraktikabel macht.[30]

EIDR basiert auf dem Digital Object Identifier (DOI)[31], der auf dem Handle-System[32] aufbaut und wurde 2010 zur Unterstützung automatisierter Workflows im Rahmen der Verwertung entwickelt.[33] Auch wenn es verstärkt Bemühungen der Registry Association um öffentliche Einrichtungen als Metadatenlieferanten gibt, sind diese in den letzten Jahren kaum sichtbar geworden.[34] Das Datenmodell von EIDR ähnelt FRBR und EN 15907, und arbeitet mit den Entitäten Abstraction (abstract work), Edit (creative version), Manifestation (technical version) und Collection (compilation). Das Parent-Element eines Basismetadatensatzes, der BaseObjectData-Block, entspricht dem Audiovisuellen Werk und enthält an Elementen eine Teilmenge von EN 15744.[35] Obwohl die EIDR-ID auf dem Handle-System basiert und somit neben dem EIDR-Portal auch über das Handle-System auflösbar bleibt, sind das Datenmodell und die einzelnen Elemente nicht in einer Data-Type-Registry[36] publiziert und damit nicht automatisiert einsetzbar. Das Fehlen einer maschinenlesbaren Beschreibung der Datentypen verlagert die Einhaltung der Schemakonformität auf die jeweiligen teilnehmenden Institutionen.

EBUCore

Um dem Bedarf der Interoperabilität zwischen Archiven Rechnung zu tragen und um selbsterklärende Archivpakete zu erzeugen, ist ein passendes Metadatenaustauschformat erforderlich. Wie bereits erwähnt, konnte sich bisher für keinen der Standards EN 15907 und EN 15744 ein Austauschformat durchsetzen. Für das Metadatenset des audiovisuellen Werkes bietet sich der Standard EN 15744 besonders an, weil er auf Dublin Core[37] beruht und die Identifizierung und Disambiguierung zum Ziel hat. Deshalb wurde durch die Projektpartner in enger Zusammenarbeit mit der European Broadcasting Union (EBU)[38] ein Mapping des semantischen

26 “Both ISAN-IA and EIDR operate as non-profit, cost-recovery organizations to the benefit of the industry”, https://www.eidr.org/eidr-isan-ia-announce-joint-registration-service/ [05.10.2021].
27 https://www.eidr.org/eidr-isan-ia-announce-joint-registration-service/ [05.10.2021].
28 ISO 15706-1:2002, Information and documentation — International Standard Audiovisual Number (ISAN) — Part 1: Audiovisual work identifier, https://www.iso.org/standard/28779.html [05.10.2021].
29 ISO 15706-2:2007, Information and documentation — International Standard Audiovisual Number (ISAN) — Part 2: Version identifier, https://www.iso.org/standard/35581.html [05.10.2021].
30 http://www.isan-deutschland.de/tarif.cfm [07.10.2021].
31 Digital Object Identifier (DOI), https://www.iso.org/standard/43506.html [07.10.2021].
32 Handle-System, https://handle.net/ [07.10.2021].
33 Bohn, *Film-Metadaten. Standards der Erschließung von Filmen mit RDA und FRBR im internationalen Vergleich und Perspektiven des Datenaustauschs*, *2018*, 27, https://doi.org/10.25969/mediarep/12912.
34 https://www.eidr.org/uam/ [05.10.2021].
35 https://www.eidr.org/documents/Introduction%20to%20the%20EIDR%20Data%20Model.pdf
36 http://typeregistry.org/registrar/#pages/About [10.12.2021]
37 Dublin Core, https://dublincore.org/about/ [10.12.2021].
38 EBU (European Broadcast Union), https://www.ebu.ch/home [10.12.2021].

Standards nach EBUCore[39] entwickelt. EBUCore ist ein interoperabler, aktiver und akzeptierter internationaler Standard für Mediendaten, der sowohl ein XML-Schema als auch eine Ontologie mitbringt. Dabei liegt der Fokus von EBUCore auf der Beschreibung der physischen (technischen) Charakteristika der Mediendaten und er ist daher auch als Zielformat für die Metadaten der AV-Datenobjekte sehr gut geeignet. Da es sich bei den filmografischen Daten um ein Minimum-Set handelt, ist dieses ebenfalls gut in EBUCore darstellbar.

Die Verwendung nur eines Standards für deskriptive und technische Informationen erleichtert die Umsetzung, spätere Auswertungen und zukünftige Kooperationen. Folglich wurde das Mapping nach EBUCore um ausgewählte Metadaten der Erfassungsebenen Manifestation und Item des Standards EN 15907 erweitert.

Auf ein vollständiges Mapping aller vier Erfassungsebenen des EN 15907 Standards in das Zielformat EBUCore wurde jedoch verzichtet. Gründe dafür sind folgende: Die konzeptionelle Ausarbeitung des drei- bzw. vierstufigen Erfassungskonzeptes ist auch nach FIAF-Richtlinien sehr volatil und von Ermessensentscheidungen des einzelnen Archivars geprägt. Besonders die Manifestations- und die Variantenebenen bieten Interpretationsspielräume und es ist mit Veränderungen in mittlerer Zukunft zu rechnen. Aus diesem Grund scheint es nicht sinnvoll, das Konzept in der Struktur der Metadaten zur Langzeitarchivierung zu verankern. Ziel war eine möglichst flache Umsetzung, die es auch für nicht drei- oder vierstufig erfassende Archive ermöglicht, die Daten zu im- oder exportieren, wobei jeweils die Zuordnung der Elemente in die eigene Systematik erfolgen kann.

In diesem Sinne wurden drei Basismetadatensets für die weiter oben definierten Entitäten Werk, Fassung und Datenobjekt bestimmt und jeweils ein Mapping nach EBUCore mit angegeben. Für das Werk gehören alle Elemente des Standards EN 15744 dazu, für Fassungen sind Angaben zum Publikations- und Produktionskontext enthalten und für Datenobjekte einige technische und deskriptive Item-Metadaten. Damit können die Metadaten eines PIDs in ein dokumentiertes XML-Format überführt werden – eine wichtige Voraussetzung für die Langzeitarchivierung und spätere Nachnutzbarkeit der Datenobjekte.

Als Teil der Langzeitarchivierungsstrategie der Deutschen Kinemathek werden nun automatisch Metadaten aus der Adlib-Datenbank extrahiert, nach EBUCore gemapped und in eine Metadatendatei im Metadata Encoding & Transmission Standard-Format (METS)[40] geschrieben, welche neben den AV-Datenobjekten in den einzelnen SIPs liegt. Da METS ein defacto Standard in der Langzeitarchivierung ist, um Strukturen und deskriptive Informationen zu erfassen, hat sich die Deutsche Kinemathek in Zusammenarbeit mit dem ZIB für das METS-Format entschieden. METS bietet die Möglichkeit, Metadaten in verschiedenen Formaten entweder direkt einzubinden, oder per referenzierter Datei zu transportieren. In beiden Fällen ist ein etablierter, dokumentierter Metadatenstandard wie EBUCore zu bevorzugen, um die Nachnutzung zu gewährleisten.
Zur Beschreibung und Strukturierung der Archivpakete wurde ein Belegungsschema des METS mit EBUCore erarbeitet, welches jeweils eigene Metadatensektionen für das Datenobjekt, die Fassung und das Werk vorsieht und mit der logischen Struktur des Pakets verknüpft.

39 EBUCore, https://tech.ebu.ch/docs/tech/tech3293.pdf [10.12.2021].
40 METS (Metadata Encoding & Transmission Standard), http://www.loc.gov/standards/mets/ [10.12.2021].

PIDs im Handle-System

Im vorangegangenen Vergleich der bestehenden PID-Systeme für Audiovisuelle Werke wurden einige Defizite erkennbar. Das sind vor allem die Ausrichtung auf den Verwertungsbereich und damit die fehlende archivarische Perspektive bei der Umsetzung, Weiterentwicklung, Vernetzung und Sichtbarmachung offener Daten. Es ist auch nicht abzusehen, wie lange die jüngst erfolgte Öffnung von EIDR gegenüber GLAM (Galleries, Libraries, Archives and Museums) und die zugesicherte unentgeltliche Nutzung der Dienste Bestand haben werden. Die Entscheidung der über öffentliche Gelder finanzierten Projektpartner fiel auf das Handle-System als ein Open-Source-Identifiersystem. “Handles”, die einzelnen Identifikatoren, bilden die technische Grundlage für den weiter oben beschriebenen DOI und EIDR. Im Unterschied zu diesen sind sie an keine weiteren Bedingungen geknüpft und gegen eine einmalige Registrierungsgebühr von US $ 50,- und eine Jahresgebühr von ebenfalls US $ 50,- einsetzbar[41]. Bei der Registrierung wird ein sogenanntes Präfix vergeben, das weltweit eindeutig die Institution kennzeichnet, hinter dem dann alphanumerische Identifier folgen.

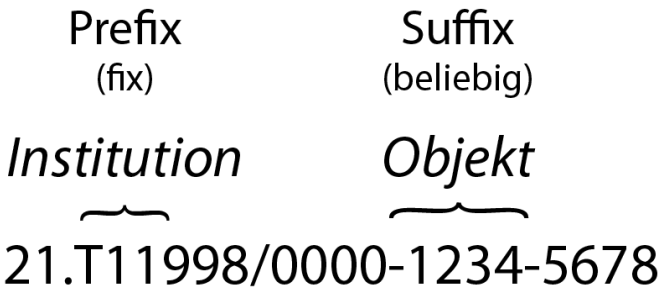

Abb. 2: Aufbau eines Handles

Da es keine vertragliche Verpflichtung gibt, den Identifikatorenservice dauerhaft zu betreiben und allein der Betrieb des Servers an Kosten gebunden ist, besteht gerade im öffentlich finanzierten Sektor die Gefahr einer nicht fortgeführten Finanzierung. Diesem aus Sicht der Persistenz nicht wünschenswerten Zustand ist das Persistent Identifier Consortium for eResearch (ePIC)[42] entgegengetreten, dessen Mitglieder sich gegenseitig garantieren, die Auflösbarkeit der Identifikatoren der jeweils anderen Institutionen sicherzustellen. Eines der Mitglieder von ePIC ist die Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG)[43], die den projektbegleitenden Testserver bereitgestellt hat und an der prototypischen Entwicklung des hier beschriebenen PID-Systems beteiligt ist.

Ein weiterer positiver Effekt der offenen Schnittstellen ist die Möglichkeit zur Nachnutzung der aggregierten Daten durch Dritte. Die Verwendung von PIDs in einem außerinstitutionellen System ermöglicht die transparente Referenzierung von auf mehrere Institutionen verteilt liegenden Datenobjekten und deren Zuordnung zu einem Werk. Die aggregierten, standardisierten Metadaten aus mehreren Archiven bieten zudem die Möglichkeit eines zentralen Verbundsystems. Dieses Ziel soll in einem Folgeprojekt umgesetzt werden und wird im Kapitel "Ausblick" ausführlich erläutert.

41 https://www.handle.net/payment.html [31.01.2022]
42 ePIC (Persistent Identifier Consortium for eResearch), https://www.pidconsortium.net [10.12.2021].
43 GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen), https://www.gwdg.de und https://www.gwdg.de/research-data-management/persistent-identifier-pid [10.12.2021].

Um eine Standardisierung zu erreichen, sind im Verlauf des Projekts in der Data-Type-Registry[44] des Handle-Systems die notwendigen Typen des EBUCore-Standards, die Basismetadaten zur Identifizierung des Werks, der Fassung und der Datenobjekte, registriert worden. Somit kann auch durch die Software anderer Institutionen, die ebenfalls persistente Handle-Identifikatoren für Ihre audiovisuellen Materialien verwenden, die Standardkonformität der Einträge geprüft werden. Im Handle-System entsteht dabei auf diese Weise ein AV-Katalog mit filmografischen Basismetadaten und einer Elementverwaltung, den Datenobjekten. Dabei wird die Zugehörigkeit der Datenobjekte zu Werk und Fassung durch die persistenten Identifikatoren im METS auch in der Langzeitarchivierung erhalten. Im Anhang ist ein technisches Beispiel einer solchen Zuordnung zu finden.

Projektergebnis

Die Implementierung von EN 15744 in EBUCore und die Umsetzung in ein METS-Format wurden zusammen mit der Publikation *Outline of an archiving strategy for digital audiovisual material*[45] als Request for Comments international und national veröffentlicht und diskutiert.[46] Die Implementierung von EN 15744 in EBUCore wurde Teil des Europäischen Standards prEN 17650 CPP, *Digital preservation of cinematographic works*[47], welcher ab 31.01.2022 zur Verfügung steht. Damit besteht die Möglichkeit, dass sich EN 15744 mit dem EBUCore Schema gegenüber Eigenentwicklungen durchsetzt.

Bei der Anbindung an die Adlib-Schnittstelle wurde ein mehrstufiger Prozess gewählt, um eine konzeptionelle Abkopplung der standardorientierten Datenhaltung vom proprietären Format der Adlib-Software zu realisieren und gleichzeitig eine Generalisierbarkeit und Wiederverwendbarkeit eben jener Softwarekomponenten zu ermöglichen, die mit den Standardformaten hantieren und in der Folge auch für andere Institutionen interessant sein werden. So wird über die Export-Schnittstelle aus Adlib ein METS-Container mit Metadaten im EBUcore Schema erzeugt. Anschließend wird dieser METS-Container über eine nachgeschaltete Softwarekomponente eingelesen, die Metadaten auf das Datenformat des Handle-Systems gemappt, welches an das Handle-System übergeben werden kann. Dabei wurde auch berücksichtigt, dass aus dem Suchergebnis des Handle-Servers eine Verlinkung in den Verleihbestand der Kinemathek möglich wird.

Durch die im Projekt entwickelten Softwarekomponenten wird es im Archivierungsprozess möglich, ein Datenobjekt der Kinemathek in ein OAIS-konformes SIP zu überführen, das mit geringen Anpassungen der METS-Datei auch von anderen Langzeitarchiven verarbeitbar ist. Wenn dieses SIP von der Ingestpipeline des EWIG Archives[48] zufriedenstellend evaluiert und in das Langzeitarchiv übernommen wurde, wird eine automatisierte Bestätigung des Transfers of Custody (der

44 DTR (Data-Type-Registry), https://dtr-pit.pidconsortium.net/# [10.12.2021].
45 Oltmanns und Hasler, *Outline of an archiving strategy for digital audiovisual material*, *2020,* ZIB Report, Zuse Institute Berlin, https://doi.org/10.12752/3.7884.
46 Zum Beispiel bei dem Projektteam CEN TC 457, Nestor AV-Media, FIAF, TIB, SLUB, Finnisches Filmarchiv, BFI, Kinemathekenverbund, Adlib-User-Gruppe und weitere.
47 prEN 17650, https://standards.iteh.ai/catalog/standards/cen/cd05d0f8-88aa-46e7-957d-28b4e638cb68/pren-17650[10.12.2021].
48 EWIG-Archiv, https://ewig.zib.de/ [10.12.2021].

Verantwortungsübernahme für die Archivalie durch das digitale Langzeitarchiv) erzeugt und an die Deutsche Kinemathek übermittelt.

Die im Rahmen des Projektes entwickelten Lösungen sollen durch konsequenten Einsatz von Open-Source-Komponenten, der Vermeidung von hochpreisigen Dienstleistern und der Offenlegung aller Schnittstellen auch kleineren filmhaltenden Institutionen zugutekommen. Die entwickelten Lösungen im Detail sind: die Implementierung der EN 15744 Metadaten in EBUCore, die Erstellung eines validierten METS-Schemas mit eingebetteten EBUCore-Sektionen für das Werk, die Fassung und das Datenobjekt und die Registrierung der EBUCore-Typen in der Data-Type-Registry. Die Software für die Extraktion der Basismetadaten aus einem METS-Format, die Umwandlung in ein JSON-Format[49] und die Registrierung einer Handles-PID ist mit geringen Abwandlungen von anderen Institutionen nachnutzbar. Auch die Extraktion von Metadaten aus Adlib, das Mapping nach EBUCore und das Schreiben eines METS-Formates sind als offene Softwarekomponenten erarbeitet worden.

Danksagung

Dieser Bericht stellt in der gebotenen Kürze die Ergebnisse unseres Projekts dar, das ohne den Einsatz der folgenden Personen weit über das erwartbare Maß hinaus nicht so erfolgreich verlaufen wäre. Auf diese Weise gedankt sei:

Sven Bingert für seine besonnene Art uns gegenüber, wenn wir wieder irrwitziges vom Handlesystem verlangten, *Jean-Pierre Evain* für sein bodenständig fundiertes Aufbrechen von EBUCore für uns, *Annette Groschke* dafür, dass sie uns immer wieder auf die Realitäten der filmographischen Archivierung zurückgeführt hat, *Niklas Hütter* für sein unerschrockenes Navigieren im Nomenklatur-Dschungel, *Jürgen Keiper* für das Aufzeigen weiterer Horizonte und nur qua Alphabet an letzter Stelle *Johannes Starlinger*, ohne den all die Konzepte Theorie geblieben wären.

Ausblick Nachfolgeprojekt

"Die wichtigste Quelle von Innovation ist der Zugang zu Wissen. Die wichtigsten Informationen sind heute maschinenlesbar."[50] Dieser Feststellung von Rafael Laguna de la Vera und Thomas Ramge folgt das nächste Projektvorhaben. Die Vernetzung standardisierter, maschinenlesbarer Metadaten über Institutionen und nationale Grenzen hinweg ist ein zentrales gesellschaftliches Ziel und wird immer dringlicher. Gerade öffentliche Institutionen haben dabei eine Vorbildwirkung, da deren Daten “öffentliches Gut”[51] sind.

49 JSON (JavaScript Object Notation), Datenformat in Textform.
50 Laguna de la Vera und Ramge, *Datenmonopole sind Diebstahl am Fortschritt, 2021,* https://www.welt.de/wirtschaft/article234462774/Mehr-Wissensfreiheit-Datenmonopole-sind-Diebstahl-am-Fortschritt.html [16.11.2021].
51 ebd.

Mittels eines Prototypen soll daher gezeigt werden, wie das Handle-System die Funktion einer institutionsübergreifenden, quelloffenen, zentralen Datenbank übernimmt. Es ermöglicht damit effizientere Strukturen, weitere Innovationen und knüpft so an die Tradition der großen Verbünde im Archiv- und Bibliotheksbereich an. Zusätzlich erweitert sich mit diesen Identifikatoren die Möglichkeit, automatisierte Prozesse bis zum eigentlichen Datenobjekt in der Langzeitarchivierung abzubilden, ein aktueller, drängender Bedarf, der die speicherintensiven AV-Daten in besonderem Maße betrifft.

Das in Vorbereitung befindliche Projektvorhaben nimmt die heterogene Situation der filmhaltenden Institutionen zum Ausgangspunkt, um eindeutige und nachhaltige Dokumentationsstrukturen für den Filmbestand in deutschen Kulturinstitutionen aufzubauen. Es handelt sich hierbei ausdrücklich nicht um umfassende filmografische Informationen, diese sind bereits vielfach verfügbar. Im Vordergrund steht die Zuordnung der Bestandsobjekte zu einem bestimmten Werk und zu einer Fassung in einem zentralen Verbundsystem. Gegenstand sind die von den deutschen Institutionen als relevant erachteten audiovisuellen Materialien aus den Beständen und dies umfasst alle Gattungen[52].

Die persistenten Identifikatoren als Klammer und Infrastruktur im Verbundsystem, in der Langzeitarchivierung und in den institutionellen Datenbanken, sind das Desiderat, um automatisierte Abläufe bis in die Langzeitarchivierung zu ermöglichen. Es soll so den Anforderungen der steigenden Datenmengen – an erster Stelle die Machinenlesbarkeit – begegnet und die Transparenz gefördert werden. Das Wissen um gleiche, oder zumindest ähnliche digitale Objekte über Institutionsgrenzen hinaus trägt unmittelbar zur Resilienz gegenüber katatstrophalen Datenverlusten in einer der Institutionen bei.

Während also zum einen die Interoperabilität und die Prozesse der Langzeitarchivierung der Mediendaten durch die persistenten Identifikatoren unterstützt werden sollen, sorgt das PID-Verbundsystem für die Sichtbarkeit der Bestände und stellt der Öffentlichkeit strukturierte Daten zur Verfügung. Dabei kommen die Anforderungen an die Identifikatoren und das zentrale System aus den Institutionen, die mit ihren Prozessen die Bestände aktiv und passiv sichern und zugänglich machen. Neu in der filmhaltenden Community ist dabei der unbedingte Anspruch an automatisierte Prozesse und offene Metadaten. Darüber hinaus ermöglicht und erfordert die fortschreitende Digitalisierung der Bestände andere Formen der Sichtbarkeit und damit Zugänglichkeit der Filme[53], wie sie durch ein Verbundsystem erreicht werden können.

Die Metadaten im prototypischen Verbundsystem können durch Synchronisationsroutinen der Institutionen ergänzt werden, und es ist möglich, Metadaten aus dem zentralen System wiederum in die Institutionsdatenbanken zu übernehmen. Dieser Prozess der gemeinsamen Katalogisierung ist eine der Grundideen der Bibliotheksverbünde und die Methoden zur Qualitätssicherung, wie etwa Deduplizierung und akkumulative Anreicherung sind bereits aus dem Bibliotheksbereich bekannt. Auf Grundlage der zu erstellenden Rechercheanforderungen aus den Archiven lassen sich Ähnlichkeitsparameter und Synchronisationsroutinen entwickeln.

52 Das meint unter anderem: Spielfilme, Dokumentarfilme, Zeitzeugeninterviews, Lehrfilme, Amateurfilme, Experimentalfilme.

53 Heftberger, *Den Prozess vom Ende her denken – Digitalisierung von Film zur Sicherung und Zugänglichmachung, 2020, https://doi.org/10.1515/abitech-2020-2004*.

Die im Verbundsystem generierten, persistenten Identifikatoren können mit nationalen und internationalen Plattformen wie EIDR, Wikidata, GND und Filmportal verknüpft werden. Die Öffnung gegenüber Diensten und Portalen ermöglicht eine Abbildung der eigenen Bestände unabhängig von nationalen Eingrenzungen ohne Verlust der institutionellen Identität. Ein Unterschied zu den genannten Portalen besteht darin, dass Werksebene, Fassung und Datenebene dargestellt und automatisiert erfasst werden. Ein Schwerpunkt des Nachfolgeprojekts wird in der Entwicklung der Rechercheoberfläche liegen. Die Datenbasis wurde als Grundlage in der erfolgreichen Kooperation der Deutschen Kinemathek und dem Zuse Institut gelegt.

31.01.2022