Langzeitarchiv für Audiowerke

Long time archive for audio works

Projektarbeit 2

Studiengang:	Master of Science in Engineering (MSE)
Autor/in:	Christoph Zimmermann
Betreuer/in:	Daniel Debrunner
Auftraggeber/in:	Schweizerische Stiftung Public Domain
Datum:	28. Juli 2016

Management Summary

This project is about the long term preservation of audio works like music, speeches etc. It was done in cooperation with the Swiss Foundation Public Domain which is operating the volunteer based Public Domain Project. This project is a digital repository for audiovisual cultural heritage to preserve it for future generations.

First an introduction into the field of digital long term preservation and the current state of this field of science is given. The presented models and obligations are fundamental to evaluate organizations, processes and system architectures for their ability to achieve the long term preservation goals. An audit according to CCSDS 652.0-M-1 was done to get a detailed insight to the fulfillment of these goals. Based on these results the requirements engineering was done. Finally a new system architecture is proposed for the long term archival storage and its associated ingest, delivery and management systems.

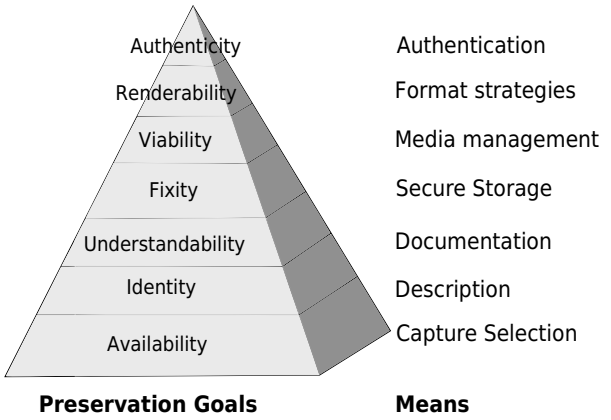

Image 1: All these properties of a digital object have to be preserved to achieve long time preservation of the information. Source: [CAP08]

Zusammenfassung

In dieser Arbeit, die zusammen mit der Schweizerischen Stiftung Public Domain durchgeführt wurde, geht es um Audiowerke wie Musik, Hörspiele, Reden etc. die digitalisiert wurden und in der digitalen Domäne langfristig erhalten werden sollen, so dass nachfolgende Generationen darauf Zugriff haben. Zuerst erfolgt eine Einführung in die Thematik der digitalen Langzeiterhaltung und das OAIS Referenzmodell.

Ein Audit gemäss CCSDS 652.0-M-1 bildet die Basis um die Erfüllung der Erhaltungsziele des Audioarchivs zu beurteilen, das von der Stiftung betreut wird. Aus den Resultaten werden Anforderungen für die Weiterentwicklung abgeleitet. Abschliessend wird eine neue Systemarchitektur für die digitale Langzeiterhaltung von Audiowerken präsentiert.

Inhaltsverzeichnis

1 Einleitung

In dieser Projektarbeit und in der daran anschliessenden Masterthesis geht es um die Entwicklung und Umsetzung einer geeigneten Systemarchitektur für ein digitales Langzeitarchiv für Audiowerke. Unter Audiowerken sind hier Aufnahmen von musikalischen Darbietungen, gesprochener Sprache (Reden, Lesungen etc.) oder Geräuschen (Vogelgesang, typische Geräuschkulisse eines Marktes etc.) zu verstehen. Die Entwicklung ist Gegenstand dieser Projektarbeit, die Umsetzung dieser Architektur wird in folgenden Arbeiten verfolgt werden.

Diese Arbeit ist in Kooperation mit der Schweizerischen Stiftung Public Domain entstanden, welche das ehrenamtliche Public Domain Projekt zur Erhaltung von gemeinfreien Tonträgern unterhält.
Für die Stiftung ist diese Projektarbeit ein wichtiger Schritt im Bestreben die Erhaltung kontinuierlich zu verbessern. Diese Arbeit und vor allem das darin enthaltene Audit sind der Beginn eines Prozesses um die nötigen Strukturen aufzubauen, die gemäss aktuellem Fachwissen und etablierten Standards nötig sind um eine Langzeiterhaltung zu ermöglichen.

Betrachtet wird in dieser Arbeit nur die digitale Langzeiterhaltung der Bestände des Public Domain Projekts. Die physische Erhaltung und die Digitalisierung der analogen Tonträger ist nicht Bestandteil dieser Arbeit.

Diese Arbeit soll in keiner Weise eine abschliessende Betrachtung oder Lösung für die Langzeiterhaltung von Audiowerken sein. Wie im weiteren erläutert wird, ist Langzeiterhaltung ein kontinuierlicher Prozess, in dem Regelmässig das technische, soziale und auch politische Umfeld des Archivs, beobachtet werden und bei Bedarf auf die veränderten Bedingungen oder Anforderungen reagiert werden muss.

Abbildung 2: Beispiel einer sehr frühen 7 Zoll Platte von Emil Berliner Records von 1896 wie sie im Public Domain Projekt auch vorhanden sind. Quelle: http://adp.library.ucsb.edu/index.php/matrix/detail/2000148104/564-Sweet_Rosie_OGrady

1.1 Struktur des Dokuments

Dieses Dokument ist unterteilt in die Bereiche:

- Einleitung, Ziele der Projektarbeit, Kontext in dem sich diese Arbeit bewegt
- Wissenschaftliche Betrachtung der digitalen Langzeiterhaltung und das OAIS Referenzmodell
- Selbstaudit nach CCSDS 652.0-M-1
- Anwendung der Resultate auf das Public Domain Projekt

1.2 Ziele der Projektarbeit

Die Ziele dieser Projektarbeit sind gemäss Aufgabenstellung:

Erstens den aktuellen Stand des Public Domain Projektes, dessen Arbeitsprozesse und digitalen Datenerhaltungsstrategie anhand von etablierten Standards in professionellen Archiven zu analysieren. Zweitens eine Strategie zu entwickeln um das Projekt auf ein professionelles Niveau zu bringen und die Anforderungen an die Softwarewerkzeuge zu schreiben die dafür nötig sind. Drittens geeignete Metadatenstandards und freie Software Werkzeuge zu evaluieren um diese Anforderungen zu erfüllen. Und viertens die schon bekannten wichtigsten Probleme anzugehen, welche sind:

- Die Metadaten der Audiodateien auf der Webseite sind nur für Menschen lesbar und sind deswegen nicht zugänglich für andere Software/Plattformen und es ist nicht möglich diese in geeigneter Weise zu exportieren, was nötig ist um die Daten in andere Softwarewerkzeuge zu Übertragen.
- Die Metadaten auf der Webseite und die Metadaten eingebettet in den Audiodateien sind nicht synchron.

Dazu müssen die derzeit eingesetzten Softwarewerkzeuge erweitert oder angepasst werden und ein abgeänderter Arbeitsprozess muss implementiert werden um diese Probleme zu lösen. Dies ermöglicht, dass die entwickelte Strategie in der folgenden Masterthesis vollständig implementiert werden kann.

2 Kontext dieser Projektarbeit

2.1 Das Public Domain Projekt und die Schweizerische Stiftung Public Domain

Das Public Domain Projekt ist ein 2009 gegründetes Projekt *zur Erhaltung und Nutzbarmachung von gemeinfreien Musik- und Filmaufzeichnungen.*[1]

In diesem Projekt engagieren sich Freiwillige dafür, Tonträger aus den Anfängen der Schallaufzeichnung zu erhalten und diese Werke im Internet für alle zugänglich zu machen. Das Projekt schlägt so eine Brücke zwischen den Anfängen der Aufzeichnung von Tönen und dem Internetzeitalter.

Damit folgt dieses Projekt vergleichbaren Projekten, die mit ehrenamtlicher Arbeit Kulturgut erhalten, allen voran ist hier das Gutenberg-Projekt[2] zu nennen, das seit 1971 gemeinfreie Bücher digitalisiert und verfügbar macht. Im Bereich Musik darf das International Music Scores Library Project (IMSLP)[3] hervorgehoben werden, das Noten von derzeit über 100 000 Werken online zur Verfügung stellt. In weiten Kreisen bekannte Projekte, die mit freiwilligen Helfern grosse Leistungen erbringen sind die freie Enzyklopädie Wikipedia oder die freie Weltkarte OpenStreetMap.

Ein Überblick zum Werdegang des Projektes ist in der Chronologie des Projektes[4] nachzulesen.

2.1.1 Schweizerische Stiftung Public Domain

Das Projekt wird getragen von der Schweizerischen Stiftung Public Domain, die bezweckt:

- *gemeinfreie Ton- und Bildaufzeichnungen, insbesondere von Musik, Filmen und Rundfunksendungen schweizerischer Herkunft, zu sammeln, zu erhalten, und in der Schweiz zu verbreiten sowie öffentlich nutzbar und bekannt zu machen.*

- *Informationen über gemeinfreie Ton- und Bildaufzeichnungen zu sammeln, zu katalogisieren, und in der Schweiz zu verbreiten sowie öffentlich nutzbar und bekannt zu machen.*[5]

Die Stiftung übernimmt die Verwaltungsaufgaben und als langfristige Institution kümmert sie sich vor allem um die analogen Tonträgersammlungen, die der Stiftung von mehreren Sammlern überschrieben wurden. So sind derzeit über 50 000 Tonträger im Besitz der Stiftung.

2.2 Was bedeutet gemeinfrei (Public Domain)

Im Namen des Projekts und der Stiftung kommt jeweils der Begriff *Public Domain* vor, es muss darum kurz darauf eingegangen werden, was darunter genau zu verstehen ist:

> *Public Domain (eigentlich «öffentlicher Grund», «Allmend») steht für diejenigen Inhalte, die nicht oder nicht mehr urheberrechtlich geschützt und damit frei verfügbar sind. Diese Inhalte sind ge-*

[1] http://de.publicdomainproject.org/index.php/PD:%C3%9Cber_PUBLIC_DOMAIN_PROJEKT
[2] http://gutenberg.org
[3] http://imslp.org
[4] http://de.publicdomainproject.org/index.php/PD:Chronologie_des_Projekts
[5] http://de.publicdomainproject.org/index.php/Statuten#Art._2_Zweck

meinfrei. Der Zugang zu Ihnen kann nicht durch das Urheberrecht begrenzt oder kostenpflichtig ausgestaltet werden.[6]

In der Schweiz sind Werke frei verfügbar, wenn der Urheber mehr als 70 Jahre tot ist und 50 Jahre seit der ersten Veröffentlichung vergangen sind. Dies ist einer der Gründe wieso sich das Public Domain Projekt auf Tonträger aus der Frühzeit der Tonaufzeichnung konzentriert. So kann das Public Domain Projekt zu den archivierten Werken uneingeschränkten Zugang bieten.

Abbildung 3: Beispiel einer Zelluloid Walze: Ansicht auf die Verpackung und die Rillen

Abbildung 4: Beispiel einer Zelluloid Walze. Ansicht auf die Stirnseite.

Quelle: Nr. 4M 1047 der U.S. Everlasting Records. Titel: My Heart Has Learned To Love You. Musik komponiert von Ernest R. Ball (1878–1927), Liedtext von Dave Reed
http://pool.publicdomainproject.org/index.php/Everlasting-4m-1047

2.3 Paradigmenwechsel von analoger Erhaltung zu digitaler Migration

2.3.1 Das analoge Dilemma

Der Wert einer Sammlung analoger Tonträger wird durch zwei Gesichtspunkte repräsentiert. Der formelle Aspekt sieht das historische Medium, etwa die Schellackplatte, deren sammlerischer Wert in ihrer Rarität besteht. Der inhaltliche Aspekt ist die Information auf diesem Medium, etwa das musikalische Kunstwerk, dessen Wert seine schöpferische Einzigartigkeit ist. Die Krux analoger Medien besteht leider darin, dass die Inhalte mit der Form zugrunde gehen. Geht die Platte zu Bruch, war es das mit der Kunst.

Analoge Medien altern und verfallen nicht allein über die Zeit, sondern auch mit jeder Nutzung. Schont man sie durch seltene Nutzung in exklusiver Runde, schadet man auch dem Inhalt, der dadurch samt seinem Schöpfer in Vergessenheit gerät. Schont man sie nicht, zerstört man jedes Mal den

[6] Antworten auf häufige Fragen Urheberrecht – Public Domain https://www.ige.ch/urheberrecht/haeufige-fragen/public-domain.html

Inhalt ein Stück mehr. Kopieren ist auch keine langfristige Lösung, denn bei einer Kopie von analog zu analog ist die Kopie aus physikalischen Gründen immer schlechter als das Original. Die Abhängigkeit von einem, im Extremfall weltweit nur einmal verfügbaren physischen Original verursacht hohe Risiken. Grosse Bedrohungen wie Naturkatastrophen (Erdbeben, Überschwemmungen, Vulkanausbrüche) oder Kriege können gleich ganze Sammlungen zerstören. Nicht zu unterschätzen sind aber auch die schleichenden Bedrohungen wie Handhabungsfehler, Diebstahl oder langsam ablaufende Zerfallsprozesse.

Dieses analoge Dilemma hat auch Folgen für die Gesellschaft, denn der kreative Schaffensprozess erfordert einen Austausch und die kritische Auseinandersetzung mit den Werken früherer Generationen. Die Erschaffung kreativer Werke ist ein Kreislauf, bestehend aus Produktion, Verteilung und Nutzung. Werke, die nicht aufgeführt, gespielt, gehört, gelesen, erlebt werden können, sind nicht mehr Teil dieses kreativen Kreislaufs, fliessen also nicht mehr in die Schaffung zukünftiger Werke ein.

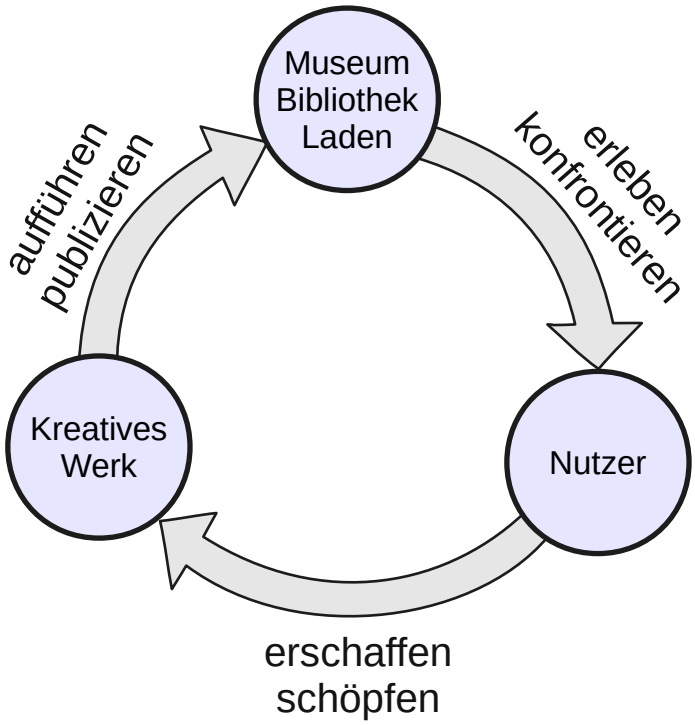

Abbildung 5: Die Erschaffung kreativer Werke ist ein Kreislauf. Werke, die nicht erlebt werden, sind nicht mehr Teil dieses kreativen Kreislaufs.

2.3.2 Digitalisierung bedeutet Trennung von Inhalt und Medium

Der einzige Ausweg aus diesem Dilemma liegt in der Ablösung des Inhalts von der vergänglichen analogen Form durch Digitalisierung der Tonträger[7]. Digitale Daten können mit wenig Aufwand beliebig oft vervielfältigt werden und an verschiedenen Orten gespeichert werden. Wir sind so auch unabhängig vom Datenträgerformat. Ein zerstörter oder durch die technische Entwicklung obsolet gewordener Datenträger kann jederzeit durch einen anderen ersetzt werden. Der kulturelle Inhalt bleibt zu 100 Prozent erhalten![8]

[7] *Die langfristige Bewahrung der primären Information eines analogen Tonträgers bedarf daher zunächst eines Transfers dieser Information in die digitale Domäne. Quelle: [TC03] Kapitel 5*

[8] Es kann nur erhalten werden, was vorher definiert wurde als zu erhaltende Inhaltsinformation (Content Information). Definition und Anforderung erfolgt in Kapitel 7.2

Die Digitalisierung eröffnet auch die Möglichkeit, die Werke über digitale Online-Portale der Allgemeinheit verfügbar zu machen. Aus Sicht der Nutzer werden digitale Archive immer stärker an Bedeutung gewinnen, da nur so ein einfacher, schneller, günstiger und barrierefreier Zugang zu künstlerischen Werken möglich ist.

2.3.3 Neue Risiken

Analoge Archive für Schallplatten und Wachswalzen haben einen nicht zu unterschätzenden Vorteil: Die mechanische Aufzeichnung der Schallwellen in Tiefen- oder Seitenschrift ist sehr einfach und deshalb gut nachzuvollziehen. Auch im Falle totalen Vergessens über die Aufzeichnungs- und Abspielsysteme und deren Parameter wie Drehzahl, Drehsinn etc. ist es anhand von wenigen Beispielen (menschliche Sprache etc.) möglich, dieses Wissen zu rekonstruieren. Es ist sogar möglich, die Tonträger ohne elektrische Energie anzuhören; dasselbe Argument wird für Mikrofilme immer wieder als Vorteil genannt, da zum Lesen nur eine Kerze und Lupe[9] nötig sei.

So einfach ist es für digitale Archive nicht. Geeignete Dateiformate zur Langzeiterhaltung sind bisher nur in wenigen Bereichen etabliert[10]. Während sich die Spezifikationen der Dateiformate noch vergleichsweise einfach überliefern lassen, ist es sehr schwierig, die Speichermedien und Ihre Abspielsysteme selbst zu dokumentieren. Falls langfristig das Wissen über die verwendete Speichertechnik verloren geht und niemand mehr die Geräte bedienen oder reparieren kann, dann ist das digitale Archiv verloren. Daher ist es unabdingbar, das benötigte Wissen an nachfolgende Generationen weiter zu geben.

2.3.4 Chancen der digitalen Langzeiterhaltung

Der grösste Vorteil digitaler Archive ist die Möglichkeit, beliebig viele Duplikate erstellen zu können, die exakt dem digitalen Original entsprechen. Technisch lässt sich sicherstellen, dass jedes Duplikat vollkommen identisch ist mit dem Original. Noch viel wichtiger: Es kann auch Jahre später überprüft werden, ob eine Veränderung stattgefunden hat. Ein digitales Archiv kann also mit geringem Aufwand an vielen Stellen weltweit gespeichert und so vor physischer Zerstörung geschützt werden. Ein aktiv gepflegtes digitales Archiv ist deshalb auch unabhängig von der Lebensdauer eines Datenträgers, weil die Datenträger regelmässig überprüft und die Daten periodisch auf eine aktuelle, zuverlässige und verbreitete Technik umkopiert (migriert) werden müssen.

Für die Allgemeinheit ist der einfache Zugriff auf die archivierten Werke massgeblich. Die Werke können uneingeschränkt genutzt werden, ohne die Langzeiterhaltung der Originale zu gefährden. Es muss sich niemand Sorgen machen, dass der Tonträger falsch behandelt wird oder sich abnutzt. Das analoge Original kann im klimatisierten Lagerraum verbleiben.

[9] *Im Notfall können die Informationen immer noch bei Kerzenlicht von blossem Auge gelesen werden.* http://www.lambdadata.ch/sicherung_mikrofilm.php

[10] Die Anforderungen an ein geeignetes Dateiformat werden in Kapitel 8.5 erläutert. Auf die aktuellen Entwicklungen für audiovisuelle Archive wird in Kapitel 10.2 eingegangen.

3 Vertrauenswürdige digitale Langzeiterhaltung

In Kapitel 2.3 wurde ein allgemeiner Überblick zur Thematik der digitalen Langzeiterhaltung gegeben. In diesem Kapitel geht es um eine vertiefte Darstellung was unter vertrauenswürdiger digitaler Langzeiterhaltung (trustworthy digital preservation) aus professioneller und wissenschaftlicher Sicht zu verstehen ist.

3.1 Begriffsdefinitionen

Wie in jedem Wissenschaftsbereich sind exakte Begriffsdefinition wichtig um eine gemeinsame Basis zu schaffen, damit unter dem selben Begriff auch das Selbe verstanden wird. Diese Definitionen dienen auch dazu abstrahierte Konzepte zu definieren und zu etablieren.

In dieser Arbeit werden durchgehend die in der deutschsprachigen Übersetzung des OAIS Referenzmodells ([NES13]) definierten Begriffe benutzt. Das OAIS Referenzmodell wird in Kapitel 4 genauer erläutert. Bei der ersten Verwendung innerhalb eines Kapitels wird jeweils der englischsprachige Begriff in Klammern dazu gestellt. Dies soll die Einordnung und den Vergleich zu internationalen Publikationen vereinfachen sowie den fachlichen Austausch fördern.

Die DIN 34644 (Information und Dokumentation – Kriterien für vertrauenswürdige digitale Langzeitarchive) verwendet zum Teil abweichende Definitionen die hier nicht verwendet werden und nicht aufgelistet werden. Eine Vergleichstabelle ist im Kommentar zur DIN 34644 ([KEI13] Seite 11) zu finden.

Deutschsprachiger Begriff	English term	Bedeutung
OAIS	OAIS	Ein OAIS ist ein Archiv, das aus einer Organisation, die Teil einer größeren Organisation sein kann, aus Menschen und Systemen besteht, das die Verantwortung übernommen hat, Information zu erhalten und sie einer vorgesehenen Zielgruppe zugänglich zu machen.
Langfristig	Long Term	Eine Zeitspanne, die lange genug andauert, um sich mit den Auswirkungen des Technologiewandels inklusive der Unterstützung von neuen Datenträgern und Datenformaten sowie einer sich verändernden vorgesehenen Zielgruppe auf die Information im OAIS auseinander zu setzen. Diese Zeitspanne reicht bis in die unbestimmte Zukunft.
Langzeiterhaltung	Long Term Preservation	Die langfristige Erhaltung von Information in einer für die vorgesehene Zielgruppe unmittelbar verstehbaren Form, und mit Evidenznachweisen, die ihre Authentizität langfristig unterstützten.
Inhaltsinformation	Content Information	**Ein Satz an Informationen, der das eigentliche Ziel der Erhaltung ist** oder der Teile der oder die komplette Information enthält. Es ist ein Informationsobjekt, das sich aus dem Inhaltsdatenobjekt und seiner Repräsentationsinformation zusammensetzt.

Authentizität	Authenticity	Das Ausmass, in dem eine Person (oder System) ein Objekt als das ansieht, was es vorgibt zu sein. Authentizität wird auf der Basis von Evidenz beurteilt.
Unmittelbar-verstehbar	Independently Understand-able	Ein Merkmal von Information, die hinreichend vollständig ist, um von der vorgesehenen Zielgruppe interpretiert, verstanden und verwendet zu werden, ohne dass diese auf spezielle, nicht weit verbreitete, Hilfsmittel, einschliesslich benannter Personen, zurückgreifen muss.
Vorgesehene Zielgruppe	Designated Community	Eine ausgewiesene Gruppe potenzieller Endnutzer, die in der Lage sein soll, einen bestimmten Satz an Informationen zu verstehen. Die vorgesehene Zielgruppe kann sich aus mehreren Benutzergruppen zusammensetzen. Eine vorgesehene Zielgruppe wird vom Archiv definiert und diese Definition kann sich über die Zeit verändern.
Grundwissen	Knowledge Base	Eine Reihe von Informationen, verinnerlicht in einer Person oder einem System, die es der Person oder dem System erlauben, empfangene Information zu verstehen.
Repräsentationsinformation	Representation Information	Die Information, die ein Datenobjekt in für Menschen aussagekräftigere Konzepte übersetzt. Ein Beispiel von Repräsentationsinformation für eine Bitsequenz, die eine FITS-Datei ist, könnte sich aus einem FITS-Standard, der das Format definiert und einem Wörterbuch, das die Bedeutung von Schlüsselbegriffen definiert, die nicht Bestandteil des Standards sind, zusammensetzen. Ein weiteres Beispiel ist JPEG Software, die benutzt wird, um eine JPEG-Datei anzuzeigen. Die JPEG-Datei als Bits anzuzeigen ist für den Menschen nicht sehr aussagekräftig, aber die Software, die ein Verständnis des JPEG-Standards verkörpert, überträgt die Bits in Pixel, die dann als Bild für die menschliche Betrachtung angezeigt werden können.
Erhaltungsmetadaten (PDI)	Preservation Description Information (PDI)	Die Information, die benötigt wird, um die Inhaltsinformation angemessen zu erhalten, und die als Provenienz, Referenz, Beständigkeit, Kontext und Information über Zugriffsrechte kategorisiert werden kann.
Archivinformationspaket (AIP)	Archival Information Package (AIP)	Ein Informationspaket, bestehend aus der Inhaltsinformation und den dazugehörigen Erhaltungsmetadaten, das innerhalb eines OAIS aufbewahrt wird.
Übergabeinformationspaket (SIP)	Submission Information Package (SIP)	Ein Informationspaket, das vom Produzenten an das OAIS geliefert wird, um es zur Konstruktion oder zur Aktualisierung eines oder mehrerer AIPs und/oder den dazugehörigen Erschliessungsinformationen zu benutzen.
Auslieferungsinformationspaket (DIP)	Dissemination Information	Ein Informationspaket, abgeleitet aus einem oder mehreren AIPs, das als Antwort auf eine Anfrage an das OAIS von dem

	Package (DIP)	Archiv an den Endnutzer gesendet wird.
Digitale Migration	Digital Migration	Der Transfer digitaler Information innerhalb des OAIS mit dem Ziel ihrer Erhaltung. Sie unterscheidet sich von Transfer im Allgemeinen in drei Punkten: • einem Fokus auf der Erhaltung des gesamten Informationsgehalts, der Erhaltung benötigt; • einer Perspektive, dass die neue archivische Erscheinung der Information ein Ersatz für die alte ist; und • dem Verständnis, dass die volle Kontrolle und Verantwortung über alle Aspekte des Transfers bei dem OAIS liegen.
Persistenz-information	Fixity Information	Die Information, welche die Mechanismen dokumentiert, die sicherstellen, dass das Inhaltsinformationsobjekt nicht unerlaubt verändert wurde. Ein Beispiel ist ein Schlüssel aus einer zyklischen Redundanzüberprüfung (CRC) für eine Datei.
Archivverbund	Federated Archives	Eine Gruppe von Archiven, die sich darauf verständigt hat, Zugriff auf ihre Bestände über eine oder mehrere gemeinsame Findmittel zu ermöglichen
Zugriffsprogramm	Access Software	Eine Art von Software, die Teile des oder den gesamten Informationsgehalt eines Informationsobjekts in für Menschen oder Systeme verstehbaren Formen präsentiert.

Tabelle 1: Die wichtigsten Begriffe für diese Projektarbeit aus dem OAIS Referenzmodell in deutsch und englisch. Aus [NES13], Seiten 8 bis 16

3.2 Was ist ein digitales Langzeitarchiv

Die DIN 31644 definiert ein digitales Langzeitarchiv so:

> *Organisation (bestehend aus Personen und technischen Systemen), die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit von Information in digitaler Form sowie die Bereitstellung für eine bestimmte Zielgruppe übernommen hat. [KEI13]*

Was der Definition im englischsprachigen OAIS Referenzmodell entspricht:

> *[...] an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. [OAIS12]*

Wie aus der Begriffsdefinition zu entnehmen ist, bezeichnet *Langzeit* (long term) keine bestimmte Zeitspanne (Fünf Jahre, drei Generationen etc.) sondern ob die „*Auswirkungen des Technologiewandels inklusive der Unterstützung von neuen Datenträgern und Datenformaten sowie einer sich verändernden vorgesehenen Zielgruppe*“ relevant sind. Diese offenere Definition greift damit viel weiter als klassische Archive/Bibliotheken und erfasst (bewusst) viel mehr Institutionen, Firmen ja sogar einzelne Abteilungen/Personen.

Daraus folgt auch: Um diese Verantwortung übernehmen zu können, ist es lohnenswert schon zum Zeitpunkt der Erschaffung von Information Überlegungen zu deren Langzeiterhaltung anzustellen.

3.3 Was bedeutet vertrauenswürdig (Trustworthy)

Damit von einer vertrauenswürdigen Langzeiterhaltung (Long Term Preservation) gesprochen werden kann, muss Vertrauen vorhanden sein sowohl gegenüber der archivierenden Institution wie auch in die Information die von dieser archivierenden Institution erhalten wird.

3.3.1 Vertrauen in die Institution

Die vorgesehenen Zielgruppen und die Geldgeber müssen das Vertrauen haben, dass die Institution der übernommenen Verantwortung gewachsen ist. Es ist die Aufgabe der Institution dies in geeigneter weise darzulegen. Dazu gehört die Arbeitsweise, die Finanzierung, die Fähigkeiten der Mitarbeitenden, die technische Ausrüstung und wie die Institution diese Aspekte kommuniziert.

Daraus ergibt sich z. B. ganz natürlich die Erwartungshaltung der vorgesehenen Zielgruppen und der Geldgeber, dass eine Institution zur Transparenz verpflichtet ist und eine weitreichende Offenlegungspflicht hat.

3.3.2 Vertrauen in die Echtheit der Information (Authenticity)

Damit die Informationen aus einem Archiv genutzt werden können, muss von der vorgesehenen Zielgruppe nachvollzogen werden können, dass die Information wirklich das ist, was sie vorgibt zu sein: Die Authentizität muss belegt werden können.

Das ist keine neue Anforderung im Digitalen, dieser Anspruch stellt sich generell an ein Archiv. Daher ist dies ein altes Thema von Archiven mit einer eigenen Unterdisziplin in den Geschichtswissenschaften namens Diplomatik. Generell ist die Anforderung, dass die Herkunft (Wer, wann, warum, wo) und der Kontext der Information Nachvollzogen werden kann und dass innerhalb des Archivs die Information nicht verändert wurde.

> *Provenienzinformation dokumentiert die Geschichte der Inhaltsinformation. Sie nennt den Ursprung oder die Quelle der Inhaltsinformation, alle Änderungen seit ihrer Entstehung und wer sie seitdem in Obhut hatte, und liefert damit eine Nachweiskette für die Inhaltsinformation. Das gibt künftigen Benutzern eine gewisse Zusage, wie zuverlässig die Inhaltsinformation ist, da es zum Nachweis der Authentizität beiträgt. [NES13] Seite 61*

Das Mass, wie detailliert diese Provenienzinformation gepflegt sein muss, ist bei jedem Archiv verschieden und hängt von den zu erhaltenden Informationen ab. Ein Archiv muss beim Übernehmen von Archivgut darauf achten, dass Provenienzinformationen mitgeliefert werden.

Der Nachweis der Authentizität von Information war innerhalb der Fachgemeinschaft der eigentliche Ursprung der Zertifizierungen und Audits.[11]

[11] [KEI13] Seite 22

3.4 Erhaltungsziele

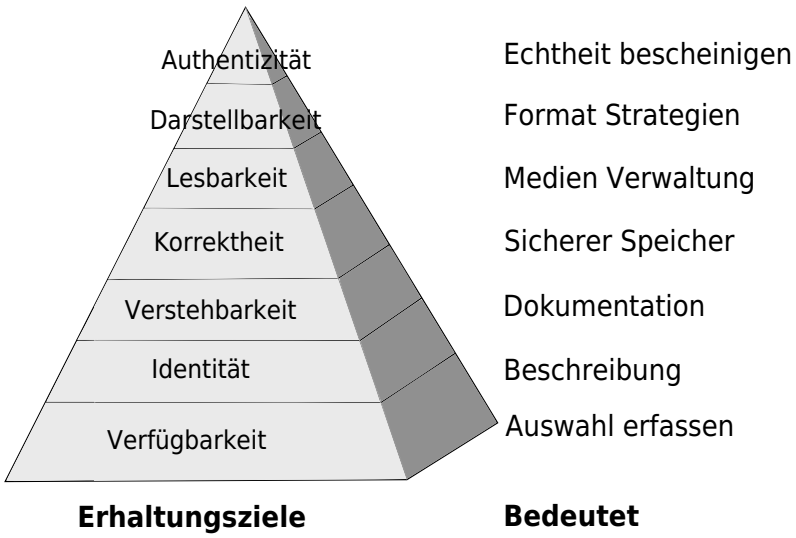

Abbildung 6: Erhaltungspyramide (Deutsche Übersetzung der "Preservation pyramide" aus [CAP08])

Aus den vorangegangenen Definitionen und als kompakte Zusammenfassung kann die *Preservation Pyramide* von Priscilla Caplan angesehen werden. Es zeigt den kompletten Umfang der Eigenschaften eines digitalen Objekts die erhalten werden müssen, damit die Information langfristig erhalten bleibt.

Erhaltungs-ziele	Preservation Goals	Bedeutung
Authentizität	Authenticity	Siehe Kapitel 3.3.2, Vertrauen in die Echtheit der Information (Authenticity)
Darstellbarkeit	renderability	Sicherstellen, dass alle benötigten Informationen verfügbar sind um eine Bitsequenz zu interpretieren und in eine aussagekräftige Form umzuwandeln.
Lesbarkeit	viability	Regelmässige Überprüfung der Datenträger ob die Datenobjekte korrekt lesbar sind. Ausfallsicherheit durch Redundanz und Backup-systeme schaffen.
Korrektheit	fixity	Sicherstellen, dass die Datenobjekte nicht unerlaubt verändert wurden.
Verstehbarkeit	understand-ability	Ein Merkmal von Information, die hinreichend vollständig ist, um von der vorgesehenen Zielgruppe interpretiert, verstanden und verwendet zu werden, ohne dass diese auf spezielle, nicht weit verbreitete Hilfsmittel, einschliesslich benannter Personen, zurückgreifen muss. [NES13]
Identität	identity	Wissen um was es sich handelt und in welchem Kontext es steht. Idealerweise ist ein digitales Objekt selbstbeschreibend.
Verfügbarkeit	availability	Ein Archiv kann nur erhalten, was im Archiv vorhanden ist und muss sich darum kümmern, dass zu erhaltende Information auch ins Archiv gelangt. Das Archiv muss das Recht besitzen Erhaltungsmass-nahmen durchzuführen.

Tabelle 2: Erhaltungsziele und deren Bedeutung

3.5 Werdegang dieses Wissenschaftsbereich

Nach der Erläuterung des derzeitigen Kenntnisstands folgt ein grober nicht vollständiger und nicht abschliessender Blick zur Geschichtlichen Entwicklung dieses Wissenschaftsbereichs.

Eine Abhandlung zum geschichtlichen Werdegang seit dem 17. Jahrhundert bis zu den Kriterien in den aktuellen Auditverfahren kann in [KEI13] Seite 20 bis 28 nachgelesen werden.

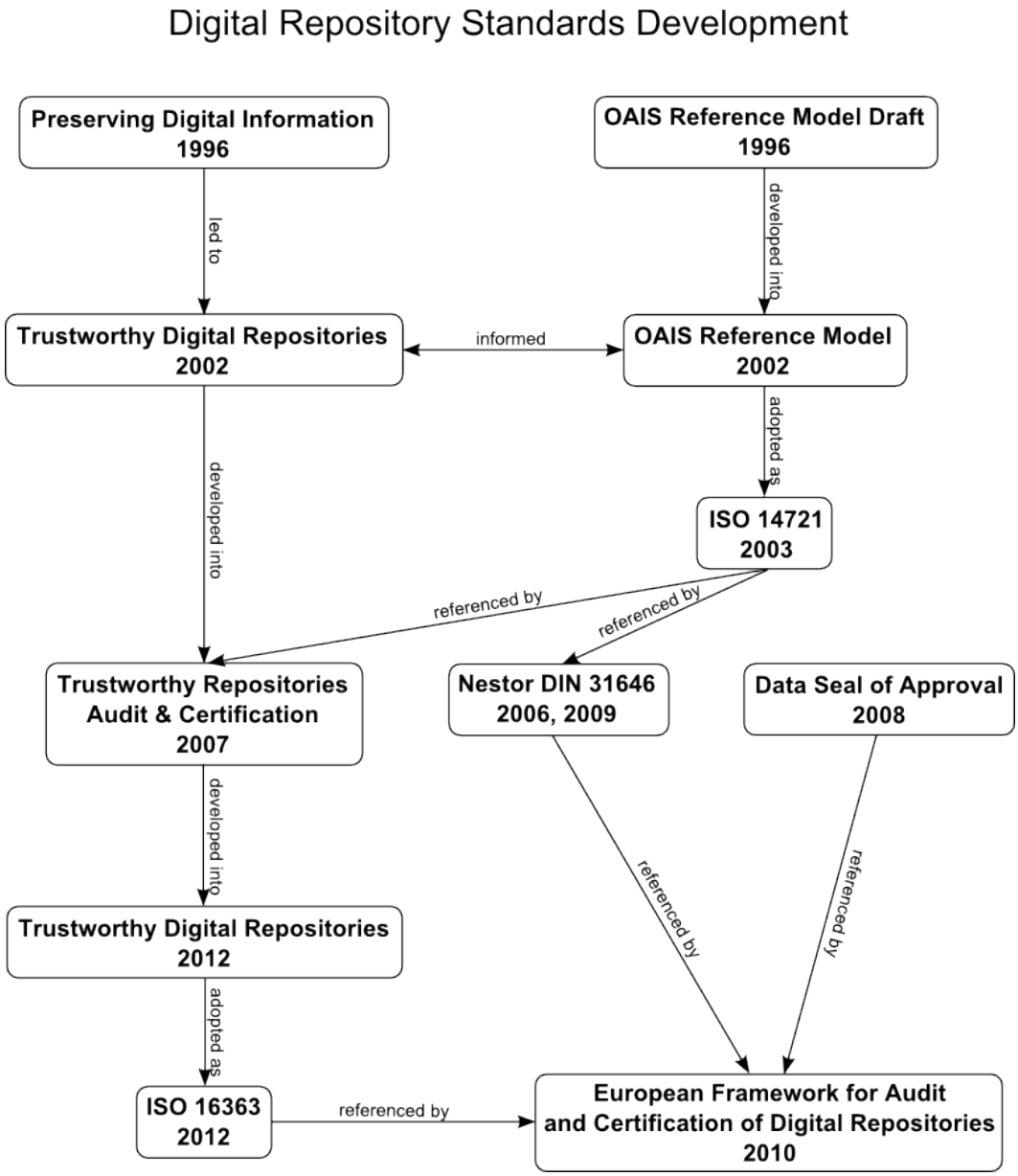

Abbildung 7: A diagram of the development of digital repository standards including OAIS (ISO 14721) and TDR (ISO 16363). Quelle: Nkrabben Wikimedia Commons (CC BY SA 3.0)

Die fachliche Diskussion beginnt schon in den 1960ern im Bereich der Astronomie und Weltraumfahrt bedingt durch die grossen Mengen an digitalen Daten, die die wissenschaftlichen Experimente liefern.[12] Daraus entstand das erste Modell zur Langzeiterhaltung: Das Goddard „Data Archive Functional Diagram“ von 1967[13]. Dieses stellte sich später als zu limitiert auf technische Herausforderungen her-

[12] Schon 1967 waren 140 000 Magnetbänder zu verwalten mit 35 000 Bändern Zuwachs pro Jahr. [KEI13] Seite 7

[13] [OADE14] Seite 10

aus. Daraus entwickelte sich dann die jetzige Definition aus dem OAIS Standard, dass ein Langzeitarchiv aus *„einer Organisatioation [...] , aus Menschen und Systemen besteht[...]““*[14].

Dieses erste Modell kannte noch keine klare Trennung von Information und Medium (Repräsentation) und auch die klare Orientierung an einer vorgesehenen Zielgruppe war noch nicht angedacht.

Das heutige Referenzmodell ist das Resultat der seit den frühen Neunzigerjahren in breiteren Kreisen intensiv geführten internationalen Fachdiskussionen zum Thema digitale Langzeitarchivierung. Massgeblich beteiligt an der Entwicklung des Referenzmodells war das CCSDS:

> *The Consultative Committee for Space Data Systems (CCSDS) was formed in 1982 by the major space agencies of the world to provide a forum for discussion of common problems in the development and operation of space data systems. Quelle: About CCSDS*

Die Entwicklung des OAIS Referenzmodells wurde von der International Standards Organisation (ISO) angestossen und dann beim CCSDS in Auftrag gegeben. Im Prinzip als Weiterentwicklung der frühen Arbeit der NASA aber von Beginn weg mit einem allgemeineren Fokus, nicht mehr fixiert auf Astronomiedaten.

In Europa hat das Kooperationsnetzwerk nestor die Aufgabe übernommen eine deutschsprachige Übersetzung des OAIS Referenzmodells anzufertigen. nestor hat sich sehr aktiv in die Entwicklung von Kriterienkatalogen und den später daraus resultierenden Zertifizierungssystemen eingebracht.

> *Mit nestor besteht ein Netzwerk, das spartenübergreifend betroffene Institutionen, kompetente Experten und aktive Projektnehmer zusammenbringt und u.a. den Austausch von Informationen, die Teilung von Aufgaben, die Entwicklung von Standards und die Nutzung von Synergieeffekten fördert.*
>
> *Dabei berücksichtigt nestor nicht allein deutsche Aktivitäten. Die Partner pflegen enge Kontakte zu entsprechenden Initiativen anderer Länder und beteiligen sich aktiv an europäischen und internationalen Initiativen und Projekten. nestor ist Mitglied in der European Alliance for Permanent Access einem Zusammenschluss von europäischen Forschungsinstitutionen, Nationalbibliotheken, wissenschaftlichen Gesellschaften, STM Publishers und der European Science Foundation. Quelle: nestor, Über uns*

Die sich zum Teil parallel entwickelnden Audit- und Zertifizierungssysteme haben einen grösseren Betrachtungswinkel als OAIS Referenzmodell. Das Referenzmodell definiert die Problemstellungen, die Verantwortlichkeiten und die dazu nötigen Begriffe, es fokussiert sich auf die Umsetzung der Langzeiterhaltung. Ein Audit betrachtet nicht nur Umsetzung sondern auch das Umfeld und legt einen stärkeren Fokus auf die Wichtigkeit der Organisation und der Menschen.

[14] [NES13] Seite 2

4 Das OAIS Referenzmodell

Das OAIS Referenzmodell ist die konzeptionelle Basis für alle Diskussionen und Abhandlungen im Bereich digitale Langzeiterhaltung (Long Term Preservation) und stellt so den wissenschaftlichen Konsens dar. Seit 2012 liegt die überarbeitete zweite Version vor mit dem Status *empfohlene Praxis* (Recommended Practice) vor. Das englischsprachige Original ist frei zugänglich als *REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)*, CCSDS 650.0-M-2 [OAIS12] und ist parallel mit dem selbem Wortlaut als ISO 14721 normiert.

Die offizielle deutschsprachige Übersetzung wurde angefertigt von der *Arbeitsgruppe -Übersetzung / Terminologie* von nestor. Diese ist frei zugänglich [NES13] und parallel im selben Wortlaut als Kommentar zur ISO 14721 publiziert [OADE14].

Dieses Kapitel soll und kann keine vollständige Zusammenfassung oder Ausführung des OAIS Referenzmodells sein. Es dient dazu, dem Leser einen Überblick über die grundlegenden Konzepte zu geben, die nötig sind um diese Projektarbeit nachvollziehen zu können.

4.1 Grundlegende Modelle

4.1.1 Das OAIS und seine Umgebung (Environment)

Ein OAIS steht nicht für sich alleine da und kann auch nicht in sich abgeschlossen betrachtet werden. Ein OAIS steht im Austausch mit seiner Umgebung und muss auf Veränderungen der Umgebung angemessen reagieren.

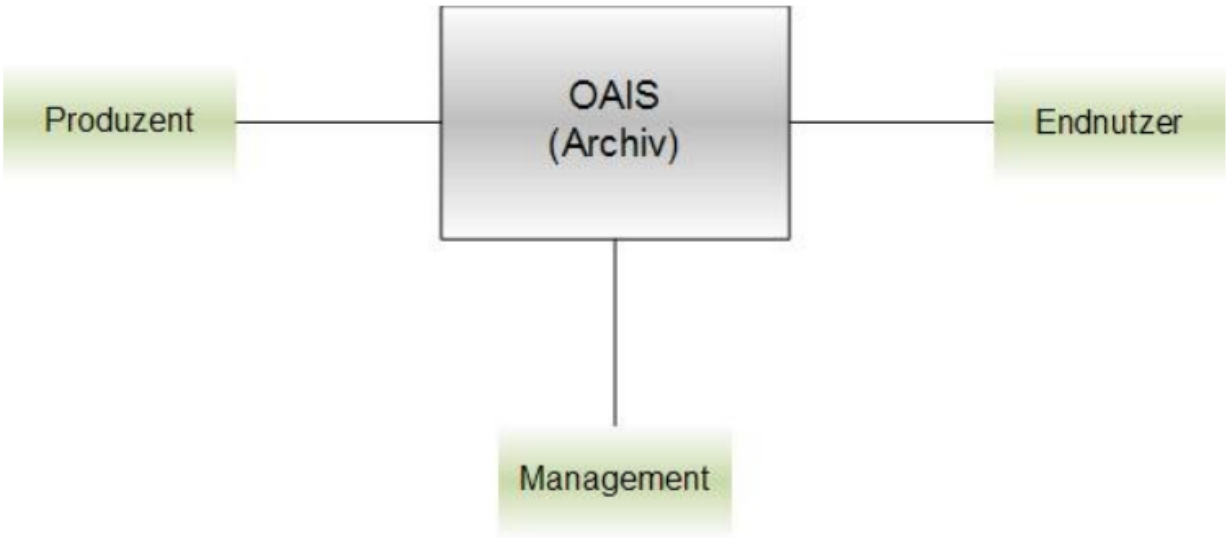

Abbildung 8: Umgebungsmodell eines OAIS [NES13], Seite 18

Endnutzer ist hier synonym zu verstehen zu allen vorgesehenen Zielgruppen.
Ein Produzent ist eine Person, Institution oder System das Informationen zur Langzeiterhaltung an das OAIS übergibt.

4.1.2 Definition von Information

> *Information (Information): Jede Art von Wissen, das ausgetauscht werden kann. Während des Austauschs wird es durch Daten repräsentiert. Ein Beispiel wäre eine Bitfolge (die Daten), beglei-*

tet von einer Beschreibung, wie die Bitfolge als Zahlen zu interpretieren ist, die eine Temperaturmessung in Celsius (die Repräsentationsinformation) darstellen. [NES13] Seite 12

Diese Definition drückt die Trennung von Informationsobjekt und Repräsentation (Expression) aus. Es ist die konsequente gedankliche Weiterführung der Trennung von Medium und Inhalt, wie in Kapitel 2.3.2 erläutert. Diese Trennung wird im digitalen weitergeführt und gehört zu den essentiellen Konzepten von OAIS die vorher auch nicht konsequent angewendet wurde. Darum stellt es eine der Haupterrungenschaften des Referenzmodells dar.

Dazu gehört das Modell, wie Information aus Daten gewonnen wird:

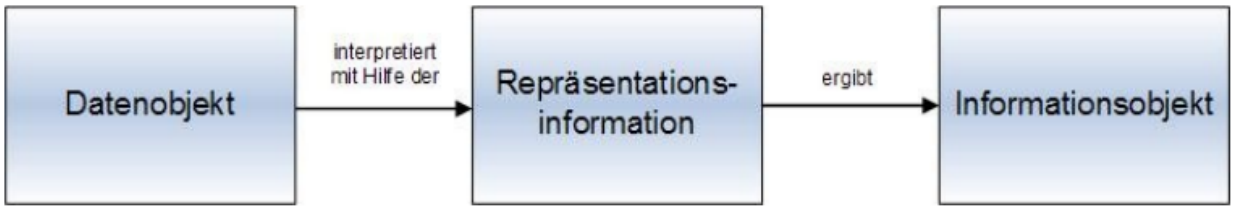

Abbildung 9: Grundlegendes Modell wie die Informationsgewinnung aus Daten erfolgt [NES13], Seite 20

Dieses Modell und der Begriff der Repräsentationsinformation ist entscheidend im OAIS Referenzmodell und für alle Langzeitarchive. Es ist die Aufgabe des Langzeitarchivs sicherzustellen, dass alle benötigte Repräsentationsinformation verfügbar ist und genau so wie das Datenobjekt erhalten wird.

Die Repräsentationsinformation kann ein ganzes Netzwerk von Informationsobjekten sein, die nötig sind um die Verstehbarkeit sicher zu stellen. Im der Abbildung 10 durch die rekursive Interpretation der Repräsentationsinformation dargestellt.

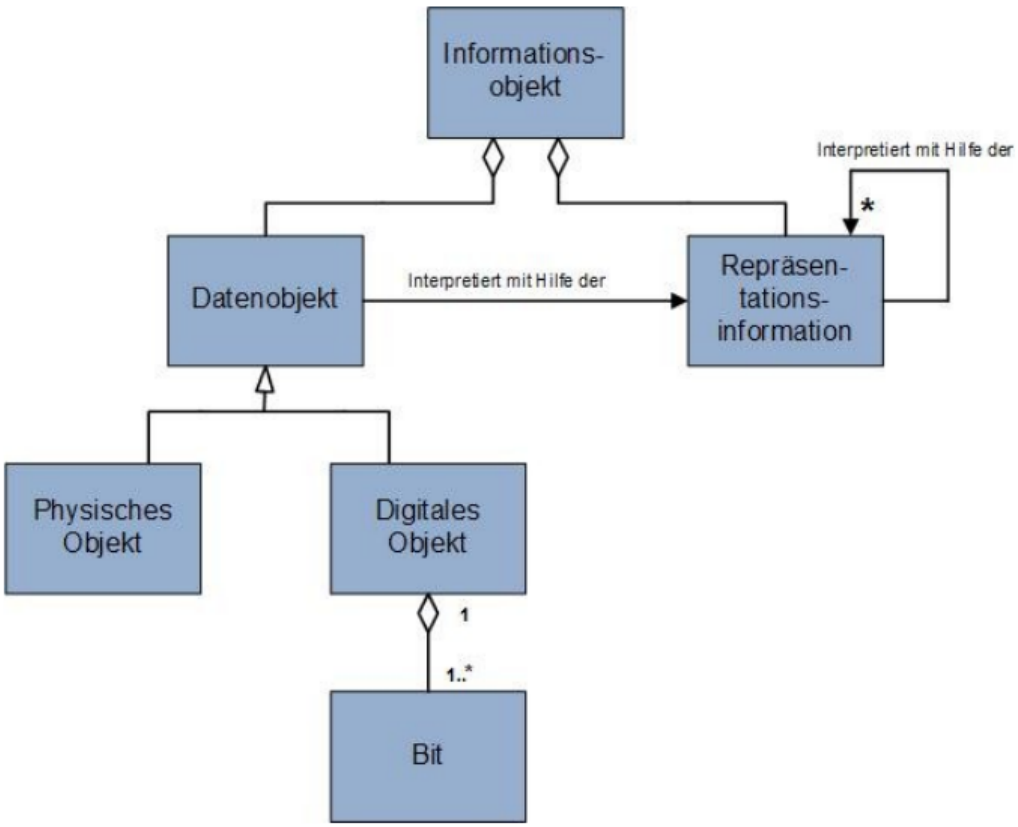

Abbildung 10: Zitat: *Das Informationsobjekt besteht aus einem Datenobjekt,das entweder physisch oder digital ist, und der Repräsentationsinformation, welche erst ein vollständiges Verständnis der Daten als bedeutungstragende Information ermöglicht.[NES13] Seite52*

4.2 Verbindliche Aufgaben gemäss OAIS

Im OAIS Referenzmodell wird definiert, was die verbindlichen Aufgaben eines OAIS sind:

> *Dieser Unterabschnitt legt die verbindlichen Aufgaben dar, die eine Organisation erfüllen muss, um ein OAIS zu betreiben.*
>
> *Das OAIS sollte:*
>
> - *mit Informations-Produzenten über Informationen verhandeln und diese entsprechend annehmen.*
> - *genügend Kontrolle über die angebotene Information bekommen, in dem Mass, das benötigt wird, um deren Langzeiterhaltung sicherzustellen.*
> - *bestimmen, entweder alleine oder zusammen mit anderen, welche Gruppen zur vorgesehenen Zielgruppe gehören sollten und deswegen fähig sein sollten, die angebotene Information zu verstehen, um dadurch ihr Grundwissen zu definieren.*
> - *sicherstellen, dass die zu erhaltende Information für die vorgesehene Zielgruppe unmittelbar verstehbar ist. Insbesondere sollte die vorgesehene Zielgruppe befähigt sein, die Information ohne den Gebrauch spezieller Hilfsmittel wie die Hilfe von Experten, die die Information erstellt haben, zu verstehen.*
> - *dokumentierten Richtlinien und Abläufen folgen, die sicherstellen, dass die Information gegen alle vorstellbaren Gefahren geschützt ist, einschliesslich der Schliessung eines Archivs, sicherstellend, dass sie niemals gelöscht wird, ausser wenn es als Bestandteil einer erprobten Strategie gestattet wird. Es sollte keine Ad-hoc Löschungen geben.*
> - *die archivierten Informationen der vorgesehen Zielgruppe verfügbar machen und die Auslieferung der Information ermöglichen, als Kopien der ursprünglich übergebenen Datenobjekten, oder zu diesen zurückverfolgbar, mit Belegen für ihre Authentizität.*
>
> *[NES13], Seite 28*

4.3 Modelle um die Auswirkungen des Technologiewandels zu bewältigen

Wie in Kapitel 3.2 gezeigt gehört zu den Kernaufgaben eines Langzeitarchivs die Bewältigung der Auswirkungen des Technologiewandels. Dies wird erreicht durch wiederkehrende digitale Migration der Informationsobjekte im Archiv. Die digitale Migration wird in [NES13] im Kapitel 5 *Perspektiven der Erhaltung* ausführlich behandelt.

Das OAIS Referenzmodell kennt im wesentlichen zwei Migrationsstrategien um diesen Auswirkungen zu begegnen. Beide Varianten unterliegen dem selben übergeordneten Ziel, die Inhaltsinformation vollständig zu bewahren. Es hängt vom zu erhaltenden Gut und dem Zeitpunkt ab, welche Strategie besser geeignet ist. Dazu müssen die Inhaltsinformation und vorgesehenen Zielgruppen definiert sein, damit die Erhaltungsplanung (Preservation Planning) danach ausgerichtet werden kann.

4.3.1 Transformation

Bei der Transformation werden die Bit-Sequenzen der Datenobjekte verändert. Typischerweise wird das verwendete Dateiformat gewechselt oder das verwendete Dateisystem auf den Datenträgern wird durch ein aktuelleres ersetzt. Die dazugehörige Repräsentationsinformation muss entsprechend angepasst werden.

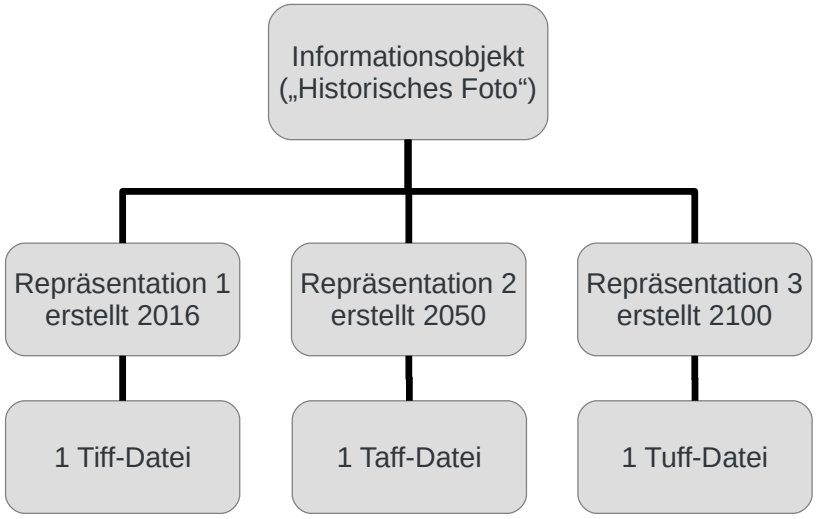

Abbildung 11: Repräsentationen bei Transformationsstrategie [KEI13] Seite 14

4.3.2 Emulation

Bei der Emulation wird die Bit-Sequenz des eigentlichen Datenobjekts nicht verändert. Stattdessen wird die Repräsentation, die aus einem Datenobjekt und einem Zugriffsprogramm besteht, im Verlauf der Zeit um Emulatoren ergänzt. Ein Emulator hat die Aufgabe, die Hardware- oder Softwarekomponente nachzubilden, die durch den Technologiewandel der vorgesehenen Zielgruppe nicht mehr zur Verfügung steht.

Ein Emulator ist im OAIS Referenzmodell eine spezielle Art von Zugriffsprogramm.

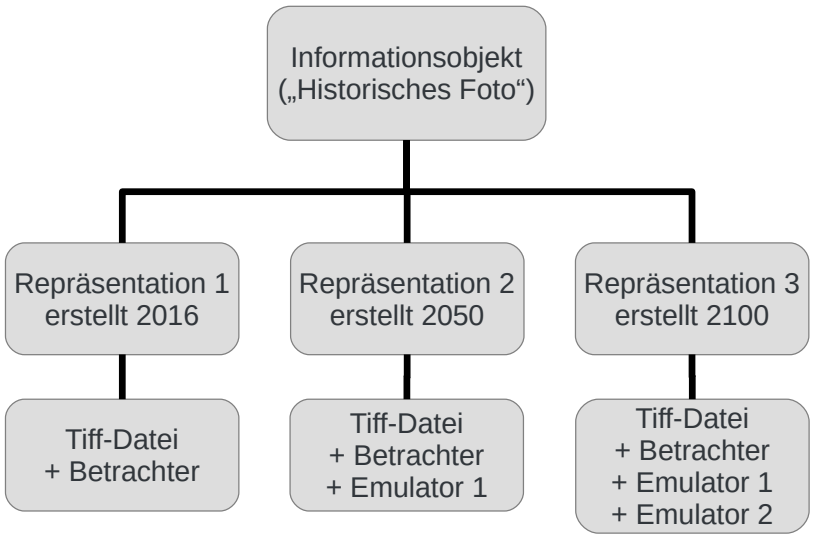

Abbildung 12: Repräsentationen bei Emulationsstrategie [KEI13] Seite 15

4.4 Funktionseinheiten eines Langzeitarchivs, Vorschlag zur Umsetzung

Das OAIS Dokument sagt ganz klar, dass:

> *Dieses Referenzmodell gibt kein bestimmtes Design und auch keine bestimmte Art der Umsetzung vor. Tatsächliche Umsetzungen können Funktionen unterschiedlich gruppieren oder herausbrechen. [NES13] Seite 3*

Das vorgeschlagene Modell besitzt zugrundeliegende Konzepte die für alle Umsetzungen wichtig sind.

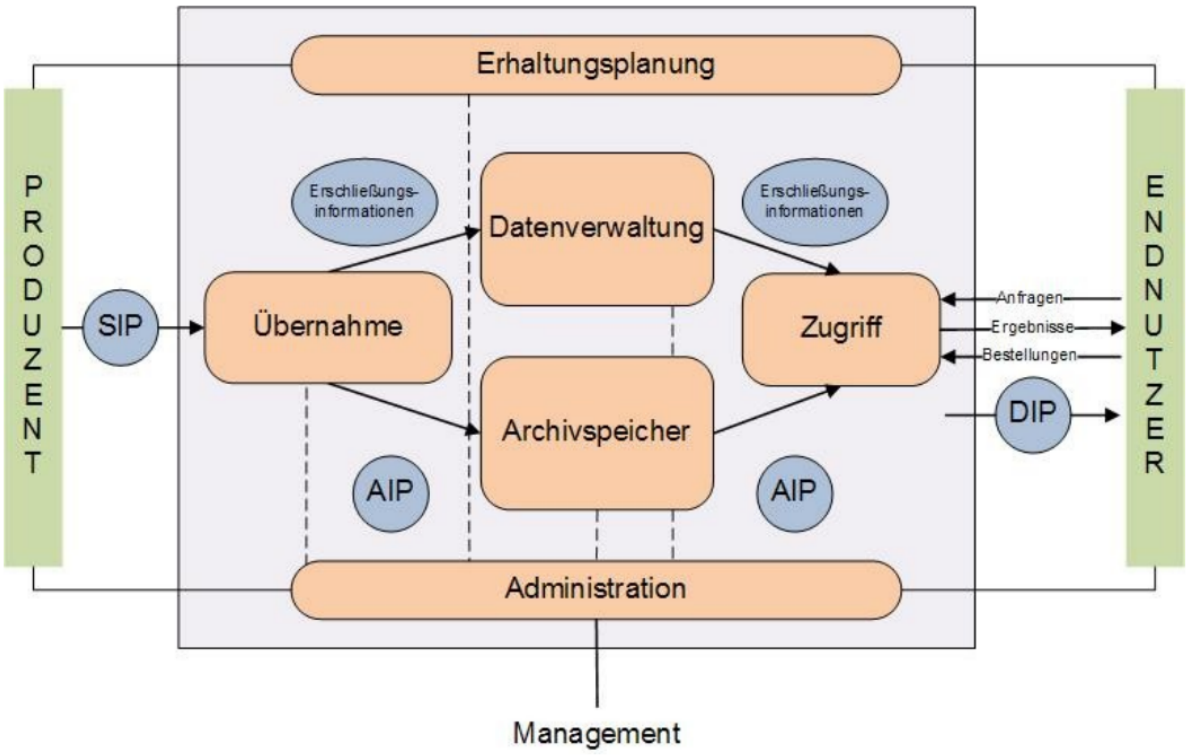

Abbildung 13: OAIS-Funktionseinheiten [NES13], Seite 33

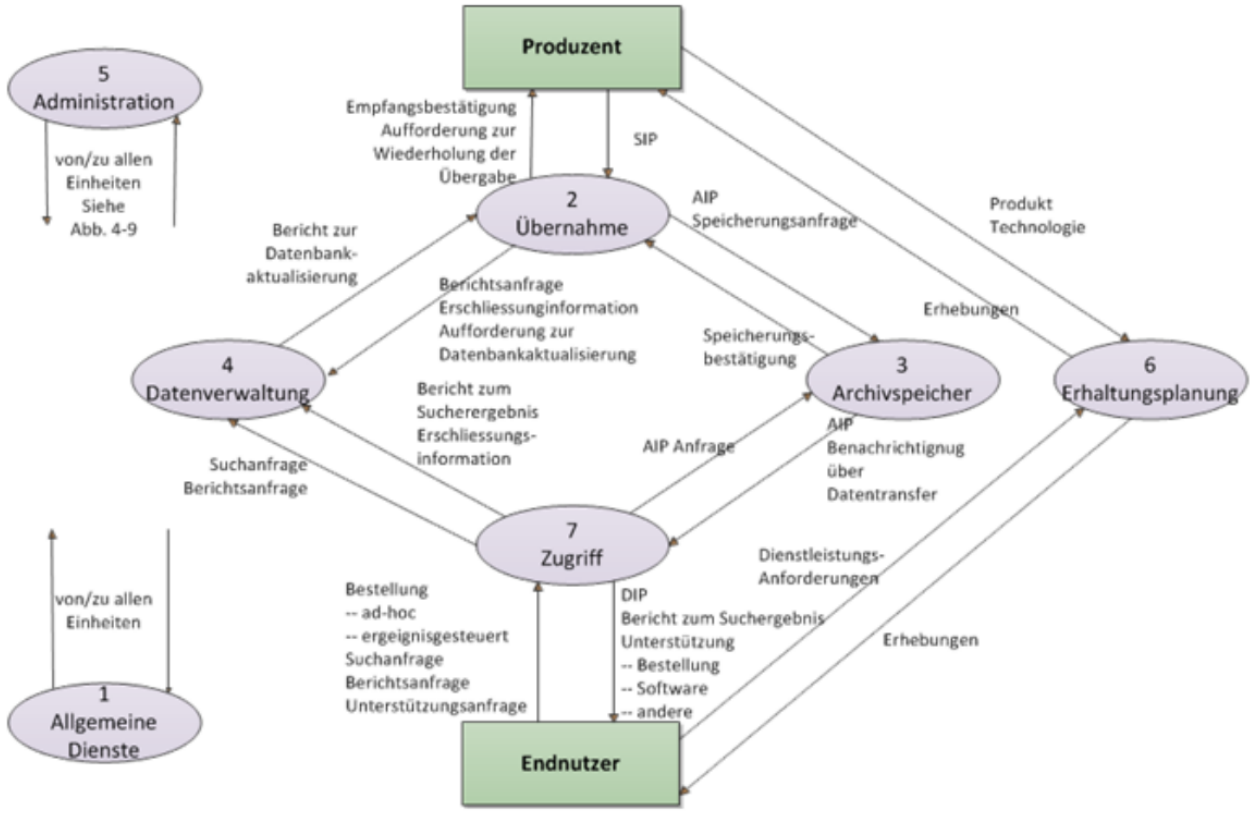

Abbildung 14: OAIS Datenfluss-Diagramm der Funktionseinheiten [NES13], Seite 50

4.4.1 Trennung von Anlieferung, Erhaltung und Auslieferung

Wesentlich an diesem Modell ist die Trennung zwischen Informationspaketen die angeliefert werden (SIP), die erhalten werden (AIP) und die ausgeliefert werden (DIP). Das bedeutet, dass all diese Objekte unterschiedlichen Formatspezifikationen folgen können und jeweils optimal auf die gestellten Anforderungen zugeschnitten werden können. In den entsprechenden Funktionseinheiten wird ein Objekt bei Bedarf in das andere Format gewandelt (Dies kann vollautomatisch ablaufen oder auch manuelle Schritte beinhalten). Diese Trennung ist für die Langzeiterhaltung sehr wichtig, denn nur so kann der Fokus auf die zu erhaltenden Eigenschaften[15] gelegt werden.

Ein kleines Beispiel zur Illustration: Der Produzent verwendet Software um seine Musikstücke zu Produzieren, die Dateien im Apple Lossless Format (ALAC) exportieren kann und entsprechend diese Dateien dem Archiv anliefern möchte. Das Langzeitarchiv wiederum nutzt zur Langzeiterhaltung eine Kombination aus WAV Dateien und XML Dateien für die Erhaltungsmetadaten. Die vorgesehene Zielgruppe wiederum möchte die Musikstücke im datenreduzierten (proprietären) AAC Format herunterladen. Ohne die angesprochene Trennung wäre es in diesem Beispiel nicht Möglich alle Teilnehmer zufrieden zu stellen und die Langzeiterhaltung sicherzustellen.

4.4.2 Aktive Erhaltungsplanung

Ein zweites wesentliches Konzept ist das Vorsehen einer Funktionseinheit zur Erhaltungsplanung. Dessen Aufgabe ist die aktive Beobachtung der Umgebung des Langzeitarchivs und soll nicht von anderen (überlasteten) Funktionseinheiten nebenbei übernommen werden. Die aktive Beobachtung geschieht durch Teilnahme an Fachdiskussionen, regelmässigen Erhebungen bei den Endnutzern und den Produzenten und Beobachtung des technischen und rechtlichen Wandels.

[15] Authentizität, Darstellbarkeit, Lesbarkeit, Korrektheit, Verstehbarkeit, Identität und Verfügbarkeit, Kapitel 3.4

Die Schweizer *Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen* (KOST-CECO) hat für den Erhaltungsprozess eine andere Darstellungsform gewählt. Vom Inhalt her entspricht es den Anforderungen und Aufgaben aus dem OAIS Referenzmodell. Darum wird hier von einer anderen Darstellungsform und nicht von einem anderen Modell gesprochen.

Die Prozessdarstellung geht von den einzelnen Arbeitsschritten aus, die nötig sind, um Information für lange Zeit zu erhalten.

> *Das Objekt des Preservation Process der KOST sind die in einem OAIS-konformen Archiv archivierten Inhalte. Der Preservation Process trägt dazu bei, diese Inhalte zugreifbar und verstehbar zu erhalten. Er umfasst damit wesentliche Teile der OAIS-Funktionseinheit Preservation Planning (ausser denjenigen, die das gesamte OAIS zum Inhalt haben), geht aber darüber hinaus, indem er auch die Umsetzung der geplanten Massnahmen umfasst. Der Gesamtprozess gliedert sich in vier Teilprozesse: Watch, Plan, Act und Check. [KOS15]*

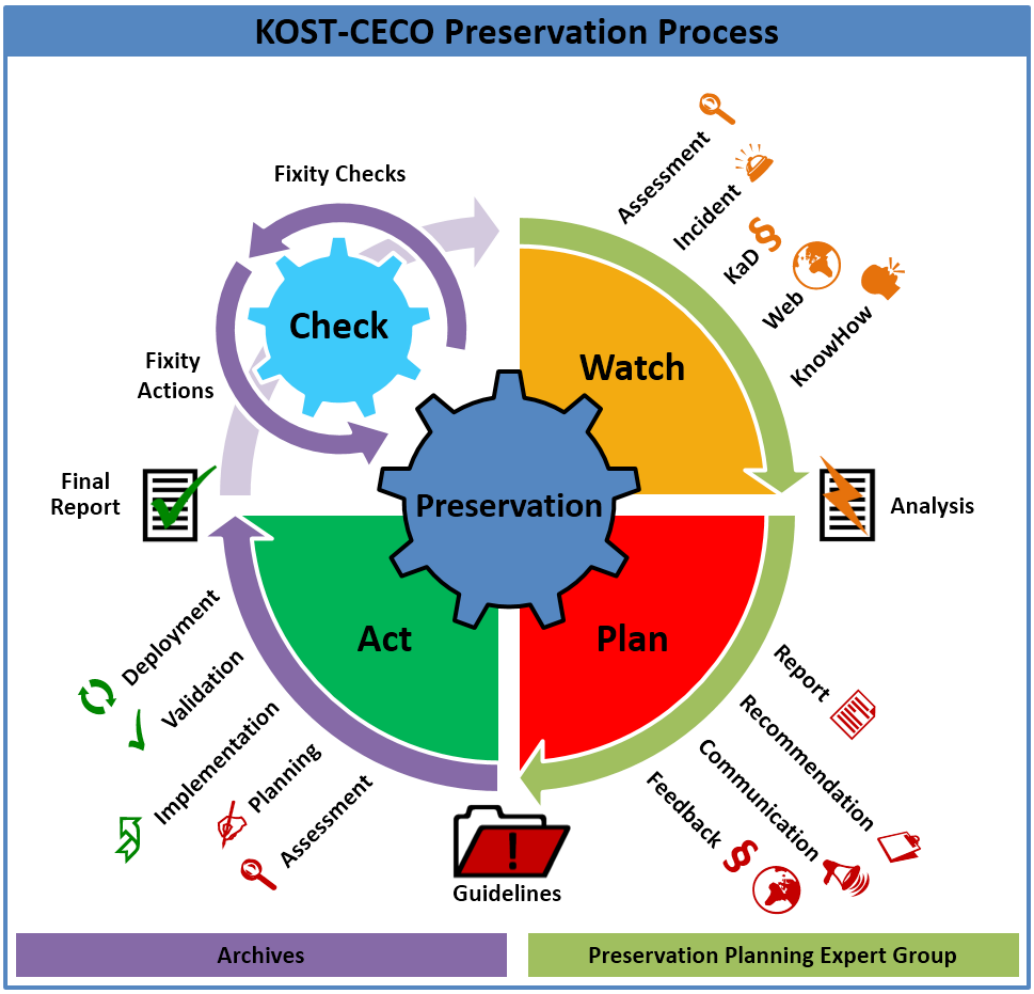

Abbildung 15: Preservation Process der Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST-CECO) [KOS15]

4.5 Klassen von Metadaten

Im Kapitel 3.4 Erhaltungsziele wurde gezeigt welche Eigenschaften erhalten werden müssen, damit ein Informationsobjekt langfristig erhalten werden kann. Ein grosser Teil dieser Eigenschaften, Darstellbarkeit, Lesbarkeit, Korrektheit, Verstehbarkeit und Identität, hängt von den Erhaltungsmetadaten (Preservation Description Information, PDI) ab.

Im OAIS Referenzmodell sind die Erhaltungsmetadaten die Vereinigung aller Metadaten, die zu einem Informationsobjekt gehören. Dabei wird zwischen verschiedenen Klassen von Metadaten unterschieden je nach engerem Zweck dieser Daten.

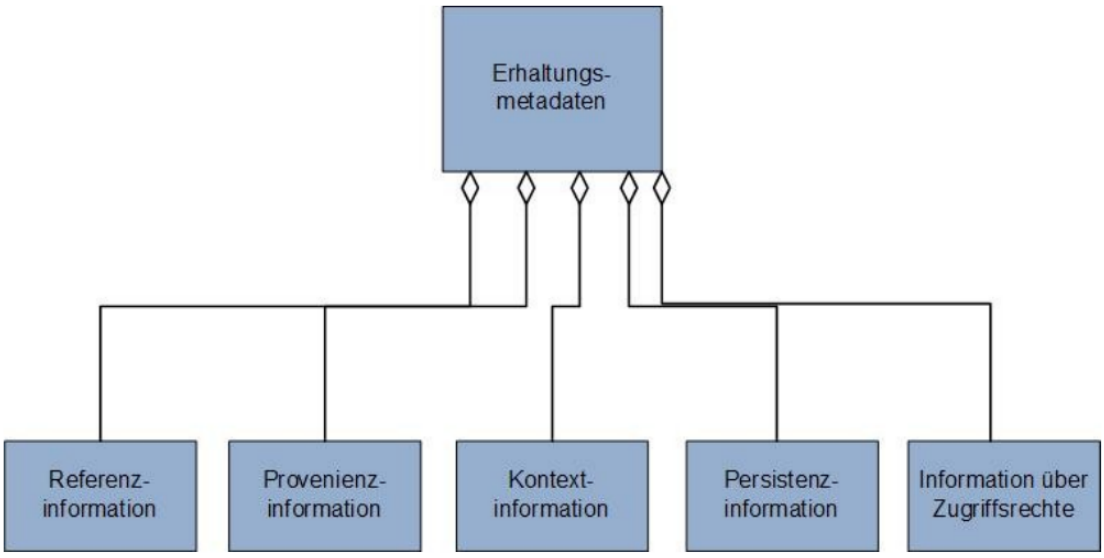

Abbildung 16: Detailmodell der Erhaltungsmetadaten [NES13], Seite 67

Klasse	Class	Bedeutung
Referenz-information	Reference Information	Eindeutiger Bezeichner um die Inhaltsinformation eindeutig zu identifizieren. Dazu gehören interne Bezeichner und auch externe Referenzbezeichner wie z. B. ISBN.
Provenienz-information	Provenance Information	Information die den Ursprung und die ganze Historie der Inhaltsinformation dokumentiert. Dazu gehört jede Änderung und wer die Inhaltsinformation in Obhut hatte.
Kontext-information	Context Information	Die Information, die die Beziehung der Inhaltsinformation zu ihrer Umgebung wiedergibt.
Persistenz-information	Fixity Information	Die Information, welche die Mechanismen dokumentiert, die sicherstellen, dass das Inhaltsinformationsobjekt nicht unerlaubt verändert wurde. Beispielsweise Prüfsummen oder Hashwerte.
Information über Zugriffsrechte	Access Rights Information	Information betreffend den rechtlichen Rahmenbedingungen, Lizenzbedingungen und Zugriffskontrolle.

Tabelle 3: Klassen von Metadaten und ihre Bedeutung

4.6 Interoperabilität von digitalen Archiven

Im OAIS Referenzmodell sind auch die möglichen Formen der Zusammenarbeit von Archiven definiert und durch entsprechende Modelle beschrieben. Aus diesen Modellen lassen sich die sehr unterschiedlichen Anforderungen ableiten, die für den jeweiligen Typ der Zusammenarbeit nötig sind.

Am Beispiel des Public Domain Projekts lassen sich zwei Typen unterscheiden:

- Kooperierende Archive (Cooperating Archives) sind Archive mit potenziell gemeinsamen Produzenten aber ohne gemeinsames Findmittel, die sich jedoch bei Standards absprechen. Die bestehende Zusammenarbeit mit Wikimedia Commons ist von diesem Typ.
- Verbundene Archive (Federated Archives) verfügen zusätzlich zur lokalen Zielgruppe über eine allgemeinere Zielgruppe die gemeinsam von den Archiven bedient wird. Dazu wird ein gemeinsames Findmittel bereitgestellt. Die Zusammenarbeit mit Memobase (CH) oder mit europeana (EU) ist von diesem Typ. Ein Bibliothekenverbund wie Nebis oder IDS sind weitere Beispiele.

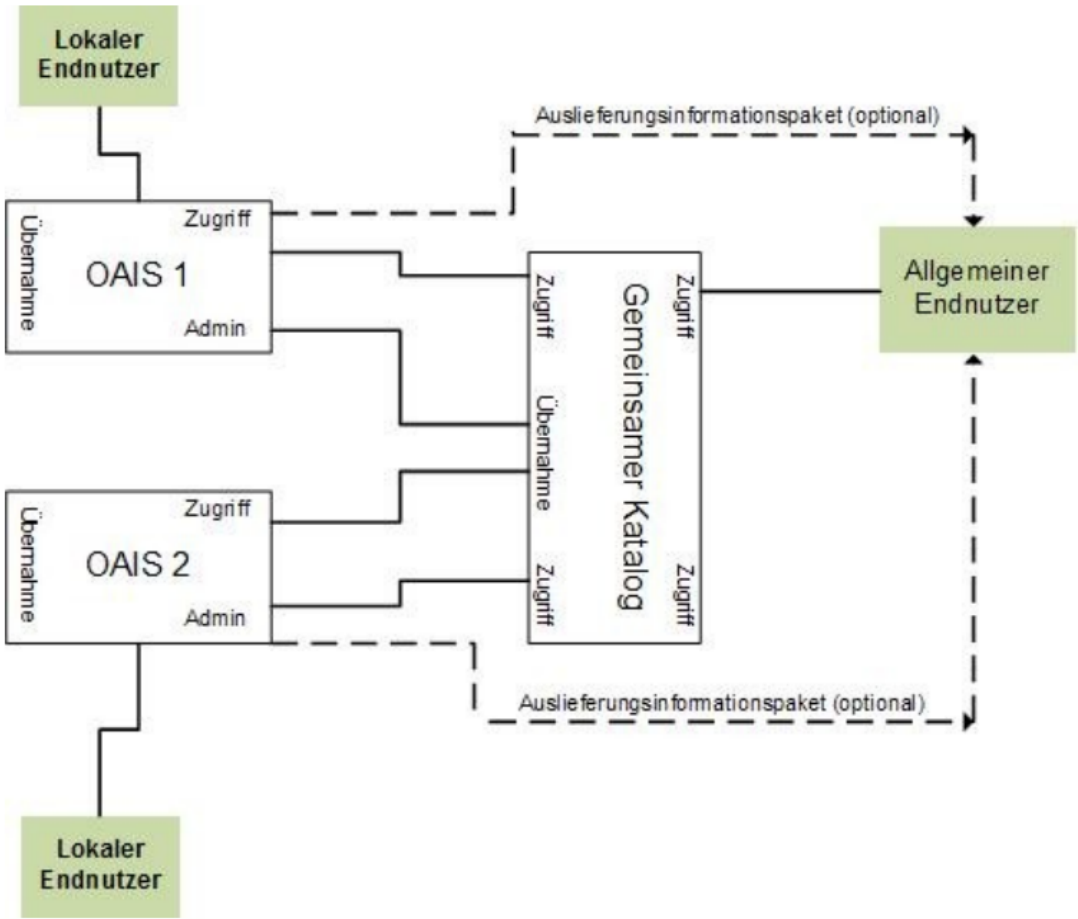

Abbildung 17: Ein OAIS-Verbund wobei ein gemeinsamer Katalog eingesetzt wird [NES13], Seite 102

Aus diesen Modellen lässt sich auch die Motivation für standardisierte Metadaten, die gegenseitig automatisiert interpretiert werden können, ableiten und der entsprechende Wunsch nach Gremien, die diese Standards koordinieren.

5 Vergleich zu aktuellen Open-* Bestrebungen

Der Begriff „Offen" in OAIS soll andeuten, dass diese Empfehlung ebenso wie zukünftige, verwandte Empfehlungen und Standards in offenen Foren entwickelt werden, und nicht, dass der Zugriff auf das Archiv unbeschränkt ist. [NES13] Seite 13

Dennoch hat ein digitales Langzeitarchiv an verschiedenen Stellen und aus unterschiedlichen Gründen Berührungspunkte zu den Open Source, Open Access und Open Data Bewegungen.

5.1 Open Source Software

Open Source Software hat auch immer zum Ziel, offene und frei zugängliche Standards zu unterstützen oder solche bereitzustellen, wenn es an solchen mangelt. Auch viele Emulatoren für obsolete Computersysteme kommen aus diesem Bereich. Das bedeutet für die Langzeiterhaltung (Long time Preservation), dass sowohl die Spezifikationen wie auch die Quellcodes der Zugriffsprogramme frei zugänglich sind und so Repräsentationsinformationen (Representation Information) bis auf Bit-Niveau erhalten werden können. Dies wird auch unterstützt durch die Anforderungen im CCSDS 652.0-M-1 Audit:

3.1.2.1 The repository shall have an appropriate succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.

Supporting Text
This is necessary in order to preserve the information content entrusted to the repository by handing it on to another custodian in the case that the repository ceases to operate.

Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement
[...] escrow of critical code, software, and metadata sufficient to enable reconstitution of the repository and its content in the event of repository failure; [...]

4.2.5.3 The repository shall have access to the requisite Representation Information. [CCSDS652]

Für die allermeisten Institutionen wird es schwierig bis kaum möglich sein mit grossen Softwareherstellern (z. B. bei Dateiformat und Dateisystemebene mit Adobe, Apple, Microsoft etc.) einen Escrow Vertrag[16] auszuhandeln. Mit dem Einsatz von Open Source Software kann die Institution selber die Verantwortung übernehmen, dass der Quellcode erhalten bleibt.

5.2 Open Hardware

Open Hardware ist für die digitale Langzeiterhaltung noch Zukunftsmusik, aber zukünftig könnten offene Datenträger mit frei zugänglicher Spezifikation in diesem Bereich entstehen. Es besteht derzeit keine Chance, genug detaillierte Repräsentationsinformationen für magnetische Festplatten, Speicherkarten oder SSDs zu bekommen. Magnetbänder oder optische Medien stehen besser da, verlieren aber an Bedeutung.

[16] Mit dem Softwareanbieter wird vereinbart, dass der Quelltext und dazugehörige Dokumentation bei einer unabhängigen Partei hinterlegt wird (z. B. Notar). Beim Eintreten bestimmter Ereignisse (vor allem bei Konkurs) wird der Quellcode dem Auftraggeber überlassen.

5.3 Open Data

Open Data und Semantic Web sind die treibenden Kräfte in der Weiterentwicklung von Metadatenstandards die nicht nur maschinenlesbar sind, sondern auch interpretierbar bleiben, wenn sie von verschiedenen Quellen kommen. Die Fachgremien der Langzeiterhaltung bringen sich in diesem Bereich aktiv ein, in dem sie Ontologien und formal definierte Vokabulare definieren (PREMIS, Kapitel 9.2.3), an der Weiterentwicklung von Ontologien mitwirken (DC Application Profile, Kapitel 9.2.2) oder an der Adaptierung bestehender Metadatenstandards an die LinkedData Konzepte mitarbeiten (FRBR).

5.4 Open Access

Open Access ist ein gutes Beispiel der sich verändernden Bedürfnisse der vorgesehen Zielgruppe (Designated Community) auch als Konsequenz der Möglichkeiten die sich durch Open Data ergeben. Bisher reichte es, wenn die Metadaten in einem Findmittel recherchiert werden konnten (Lokal oder in einem Archivverbund), neu ist der Anspruch, dass Maschinen auf Metadaten zugreifen können und diese Daten frei und ohne Lizenzeinschränkungen weiterverwendet werden dürfen. Als ideal wird angesehen, wenn Daten unter der Creative Commons CC-0 Lizenz (Keine Rechte vorbehalten, no rights reserved) vorliegen:

> *While Linked Data can be used internally within an institution or across a collaborative group, it becomes much more valuable when it is Linked **Open** Data, and is publicly shared using an open license such as the Creative Commons CC-BY or CC0 licenses*
> *Quelle: Linked Data for Libraries https://www.ld4l.org/linked-data*

Ein aktuelles Bewertungsschema der Open Data Bewegung für öffentlich zugängliche Daten ist:

★	*stelle deine Daten im Web unter einer offenen Lizenz bereit. Das Format ist dabei egal (OL = Open License)*
★★	*stelle Daten in einem strukturierten Format bereit (z. B. Excel anstelle eines einge sannten Bildes einer Tabelle) (RE = Regular Expression)*
★★★	*verwende offene, nicht proprietäre Formate (z. B. CSV statt Excel) (OF = Open Format)*
★★★★	*verwende URIs um Dinge zu bezeichnen, damit deine Daten verlinkt werden können (URI = Uniform Resource Identifier)*
★★★★★	*verlinke deine Daten mit anderen Daten um Kontexte herzustellen (LD = Linked Data) Quelle: http://5stardata.info/de/*

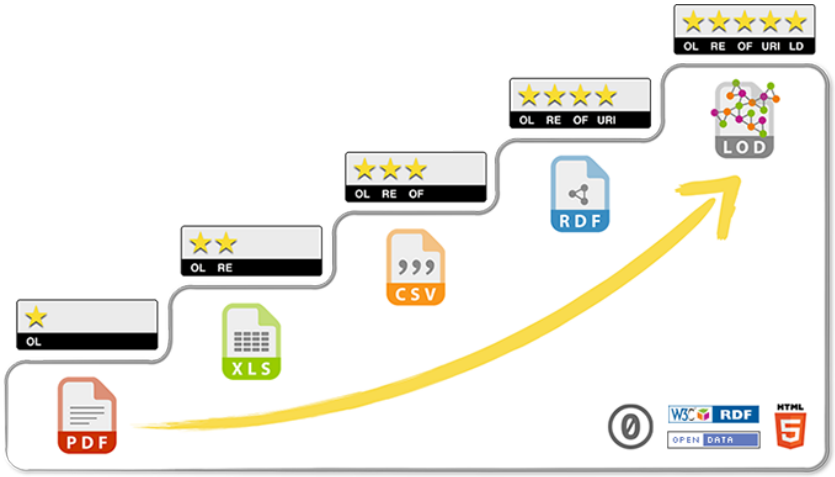

Abbildung 18: 5-Sterne-Modell für Offene Daten (*Open Data*), CC0, http://5stardata.info/de/

6 Audit nach CCSDS 652.0-M-1

Um ein Unternehmen oder ein Projekt zu verbessern, ist es nötig, zuerst den aktuellen Stand zu erfassen und diesen gemäss den definierten oder im Arbeitsbereich etablierten Zielen und Kriterien zu bewerten. Das kann mit einfachen Checklisten erfolgen oder mit aufwändigen Zertifizierungsprozessen, bei denen zum Schluss ein akkreditierter externer Prüfer die erfolgreiche Erfüllung kontrolliert und mit Zertifikat bescheinigt. Ein Audit ist eine Form von Einsatz externem Wissens und Erfahrung. Ein Audit bietet aber auch eine gute Möglichkeit, den Vorschritt zu überprüfen und die Wirksamkeit von getroffenen Massnahmen zu evaluieren.

Dieses Audit ist das erste Mal, dass innerhalb des Public Domain Projekts ein Kriterienkatalog systematisch beantwortet wird und stellt so den ersten Schritt dar um sich mit den Fragen und Anforderungen zu beschäftigen. Aufgrund der verfügbaren Informationen auf der Projektwebseite und Gesprächen mit Projektteilnehmern, ist zu erwarten, dass ein Grossteil der Anforderungen noch nicht erfüllt sein wird.

6.1 Vergleich zu anderen Auditsystemen

Wie in der Abbildung 7 zu sehen war, gab es verschiedene Bestrebungen für eine Zertifizierung von digitalen Langzeitarchiven, die aber alle eine enge Verwandtschaft haben. Hintergrundinformationen zur Entstehung und zum Werdegang dieser Auditsysteme, die parallel zur Entwicklung des OAIS Referenzmodells entstanden sind, können in [KEI13] Kapitel 2.3 und 2.4 nachgelesen werden.

Das *Trustworthy Repositories Audit & Certification [TRAC07]* gehört zu den ältesten. Parallel wurde in Europa von nestor ein längerer Kriterienkatalog, [NES08], ausgearbeitet, der dann als Basis für die Zertifizierungsbestrebungen der CCSDS und DIN diente. Mittlerweile liegen alle Anforderungskataloge in überarbeiteten Versionen vor und die grundlegenden Anforderungen wurden harmonisiert.

Zu einer anderen Kategorie gehören die Kataloge mit Handlungsempfehlungen wie z. B. der IASA [TC03] oder die *Empfehlungen zur Erhaltung von Tondokumenten* [MEM14] von Memoriav. Diese geben wichtige praktische Ratschläge, die sich konkret auf bestimmte Medien, Dateiformate, Metadatenstandards etc. beziehen.

Zur Auswahl stand auch die DIN 31644 *Kriterien für vertrauenswürdige digitale Langzeitarchive [KEI13]*. Dagegen sprach vor allem, dass darin wesentlich weniger Fragen existieren, es werden 34 Kriterien definiert, die aber umfangreicher zu beantworten sind. Es wurde darum angenommen, dass dieses Audit ohne Schulung kaum intern anwendbar ist. Es müsste ein genaueres Verständnis bestehen, was hinter den Kriterien und den verwendeten Begriffen exakt steht.

6.2 Inhalt

Das Audit nach CCSDS 652.0-M-1 *Audit and certification of trustworthy digital repositories [CCSDS652]* enthält 108 Kriterien (metrics), die in drei Hauptbereiche unterteilt sind:

Der erste Bereich betrifft die organisatorische Infrastruktur (Organizational Infrastructure) mit Kriterien zu Grundsätzen der Organisation, der strategischen Planung, Reglementen, Nachvollziehbarkeit der Entscheide, Nachfolgeregelung, Mitarbeiter und deren Qualifikation, der Finanzierung und Finanzplanung.

Der zweite Bereich betrifft die Handhabung der digitalen Objekte (Digital Object Management) mit Kriterien zur Definition der Inhaltsinformation, der Definition der Anlieferung und Prüfung auf Vollständigkeit, der AIP Definition, der Prozessbeschreibung wie aus einem SIP ein AIP wird, zum Vorhandensein der Respräsentationsinformation, der Speicherverwaltung, der Metadaten Erstellung und Handhabung und der Auffindbarkeit für die vorgesehene Zielgruppe.

Der dritte Bereich betrifft das Risikomanagement (Infrastructure And Security Risk Management) mit Kriterien zu Risikoevaluationen der potenziellen Risiken für das Langzeitarchiv, dem Life-Cycle Management der Software und Hardware, zur Sicherstellung der Korrektheit der Datenobjekte, der Computersicherheit, dem Wissen um kritische Prozesse und dem Krisenmanagement.

6.3 Vorgehen

Damit das Audit und dessen Resultate eine grosse Aussagekraft hat, wurde versucht eine möglichst neutrale und kritische Haltung zu bewahren. So wie mit einem externen Prüfer der ein Audit durchführt, wurden keine Anpassungen/Korrekturen während des Audits vorgenommen. Des weiteren werden keine Empfehlungen abgegeben sondern es wird streng nur erfasst ob die Kriterien erfüllt sind. Dies soll sicherstellen, dass der Wert des Audits für daraus folgende Weiterentwicklungen, Definitionen und Anpassungen möglichst hoch ist.

In einem ersten Schritt wurde das CCSDS 652.0-M-1 Dokument in Wikisyntax umgewandelt und als referenziertes Zitat online gestellt[17]. Damit war es möglich, im eigentlichen Bericht des Audits auf die jeweiligen Kriterien im CCSDS 652.0-M-1 Dokument zu verlinken. Damit soll die Les- und Nachvollziehbarkeit vereinfacht werden. Im zweiten Schritt wurde dann eine Vorlage erstellt, mit allen Kriterien und der erwähnten Verlinkung aber noch unbeantwortet[18].

6.4 Bewertung

Jedes Kriterium wurde nach dem Erfüllungsgrad bewertet. Dazu wurde ein Ampelsystem definiert:

Requirements fulfilled (green)	Vollständig erfüllt
Minor requirements are not fulfilled (orange)	Geringfügiges nicht erfüllt
Essential requirements not fulfilled (red)	Essentielles nicht erfüllt

Es gelten folgende Regeln für die Bewertung:

- Grün wenn ein Kriterium klar erfüllt ist
- Orange wenn ein Kriterium in wesentlichen Punkten erfüllt ist (Die Intention hinter der Anforderung ist prinzipiell erfüllt)
- Im Zweifelsfall orange statt grün bzw. rot statt orange.

Damit soll die Navigation innerhalb des Audits vereinfacht werden.

6.5 Bericht des Audits

Der Bericht des Audits ist öffentlich einsehbar, so wie es für solche Dokumente vom Audit gefordert wird. Denn für die vorgesehene Zielgruppe, die Produzenten und Geldgeber, sind diese Dokumente wichtig um Vertrauen in das Langzeitarchiv aufzubauen.

[17] http://en.publicdomainproject.org/index.php/PD:CCSDS_652.0-M-1_AUDIT_AND_CERTIFICATION_OF_TRUSTWORTHY_DIGITAL_REPOSITORIES

[18] http://en.publicdomainproject.org/index.php/PD:CCSDS_652.0-M-1_audit_template

Das vollständige Audit kann unter folgender Adresse abgerufen werden:
http://en.publicdomainproject.org/index.php/PD:Internal_CCSDS_652.0-M-1_audit

Es ist zusätzlich im Anhang dieses Berichts zu finden.

6.6 Zusammenfassung der Resultate

Die einzelnen Problemfelder die im Fazit des Audits angesprochen werden, sind hier als Übersetzung wiedergegeben.

6.6.1 Überblick

Von den 108 Kriterien wurden:

- 16 Vollständig erfüllt (grün)
- 15 Geringfügiges nicht erfüllt (orange)
- 77 Essentielles nicht erfüllt (rot)

6.6.2 Grundlegende Definitionen

Innerhalb des Public Domain Projekts und zwischen den Projektmitgliedern herrscht ein gemeinsames Verständnis der vorgesehenen Zielgruppen, diese sind aber nicht präzise definiert. Dies führt dazu, dass das Grundwissen der Zielgruppen nicht bekannt ist.

Das selbe gilt für die Definition der zu erhaltenden Inhaltsinformation.

6.6.3 Repräsentationsinformation

Ein unterentwickeltes Teilgebiet sind die Repräsentationsinformationen, da das Bewusstsein für die Notwendigkeit von Repräsentationsinformationen und der zugrunde liegenden Probleme und Konzepte vor dem Audit fehlten. Der konsequente Einsatz von offenen Standards und freier Software entspannt diese Situation ein wenig. Weil aber jegliche Repräsentationsinformation fehlt, führt das trotzdem zu einem grossen langfristigen Risiko für das Archiv.

Dieses Thema muss in der nahen Zukunft angegangen werden.

6.6.4 Verwaltung und Erhaltungsplanung

Das Gebiet der Verwaltungsaufgaben, strategische Planung, Entwicklung von Richtlinien und Verfolgung der Arbeiten ist auch unterentwickelt. Dazu kommt, dass keine Risikobeurteilung existiert und entsprechend keine Prozesse zur regelmässigen Beurteilung des technischen und rechtlichen Umfelds des Archivs existieren.

Ein System zur Unterstützung der Planung, Verwaltung und Verfolgung von Arbeiten, Meilensteinen, Fehlermeldungen etc. existiert bis jetzt nicht. Dies würde die weitere Entwicklung der Verwaltung und der Erhaltungsplanung vereinfachen.

Des weiteren fehlt ein System für Endnutzer um Rückmeldungen zu Wünschen und Fehlern einzureichen und wo sie die Reaktionen und Massnahmen zu ihrer Rückmeldung mit verfolgen können.

6.6.5 Handhabung der digitalen Objekte

Es war bekannt, dass bei der Handhabung der digitalen Objekte im Archiv Risiken existieren, die noch nicht adressiert wurden.

Die digitalen Objekte sind gefährdet weil kein System existiert, das ein ungewolltes Löschen eines Objektes verhindert, weil kein räumlich getrennte Sicherheitskopie existiert und weil kein System und die dazugehörige Überwachung existieren, um die Korrektheit auf Bit-Niveau der digitalen Objekte jetzt und in Zukunft zu garantieren.

Das System um Identifikatoren für AIPs zu erstellen ist nicht dokumentiert und ist nicht gut genug für das weitere Wachstum des Archivs.

6.6.6 Fazit des Audits

Wie erwartet, sind viele Kriterien des Audits nicht erfüllt. Trotzdem ist der Wert des Audits hoch, weil sehr unterentwickelte Bereiche innerhalb des Projekts festgestellt wurden und entsprechend das Bewusstsein für diese Probleme geschaffen wurde. Jedoch können zwölf Kriterien einfach erfüllt werden indem der aktuelle Stand dokumentiert wird.

Von hoher Priorität ist das erstellen der fehlenden grundlegenden Definitionen der zu erhaltenden Inhaltsinformation und der vorgesehenen Zielgruppen.

Ein grosser Bereich für Verbesserungen sind die regelmässigen Wartungs- und Beobachtungsaufgaben der Verwaltung und der Erhaltungsplanung. Diese müssen definiert, dokumentiert, ausgeführt und überprüft werden.

Auf der technischen Seite ist das komplette Fehlen der Repräsentationsinformation ein massives Versäumnis für ein Langzeitarchiv. Wenn diese Information in der nahen Zukunft erfasst und danach gepflegt wird, ist das reelle Risiko relativ gering, dass die Verstehbarkeit verloren geht.

7 Neu erarbeitete Definitionen

Wie dem Audit zu entnehmen ist, fehlen im Public Domain Projekt zum Teil grundlegende Definitionen bzw. sind nicht explizit ausformuliert. Diese sind essentiell um das Referenzmodell anwenden zu können und um den Erfüllungsgrad der Erhaltungsziele zu ermitteln. Entsprechend haben diese wichtigen Definitionen Auswirkungen auf alle weiteren Anforderungen. Aus diesem Grund sind diese Definitionen die ersten Resultate dieser Arbeit nach dem Audit.

7.1 Vorgesehene Zielgruppe (Designated Communities)

7.1.1 Anforderung

Die Vorgesehene Zielgruppe(n) und das von ihr verinnerlichte Grundwissen muss definiert werden. Das bildet die Basis um zu bestimmen welche Repräsentationsinformationen und Zugriffshilfen bereitgestellt werden müssen und in welcher Form diese bereitgestellt werden.

Beispiele aus dem CCSDS 652.0-M-1 Audit sind:

- *General English-reading public educated to high school and above, with access to a Web Browser (HTML 4.0 capable).*
- *For Geographic Information System (GIS) data: GIS researchers—undergraduates and above—having an understanding of the concepts of Geographic data and having access to current (2005, USA) GIS tools/computer software, e.g., ArcInfo (2005).*
- *Astronomer (undergraduate and above) with access to Flexible Image Transport System (FITS) software such as FITSIO, familiar with astronomical spectrographic instruments.*
- *Student of Middle English with an understanding of Text Encoding Initiative (TEI) encoding and access to an XML rendering environment.[...] [CCSDS652] Seite 3-6*

7.1.2 Definition

Das Public Domain Projekt hat folgende vorgesehenen Zielgruppen:

- Allgemeine Nutzergruppe (Global Community) mit Zugang zu einem Web Browser, HTML 4.0 fähig, Realschulabschluss oder höher, Sprachniveau für Englisch: A2
- Musikwissenschaftler, Historiker, Interpretationsforscher mit Zugang zu einem Web Browser, HTML 4.0 fähig, Schulabschluss: Abitur oder vergleichbar, Grundkenntnisse von DublinCore, Sprachniveau für Englisch: B2
- Suchmaschinen, Metaarchive[19], Datenanalyseprogramme (Bots) die Abfragen per HTTP 1.1 stellen können und als Antwort HTML 4.0 oder RDF 1.1 (Serialisiert als RDF/XML) akzeptieren.

[19] Zugriff über gemeinsame Findmittel, z. B. per Wikimedia Common, Memoriav Memobase oder europeana

7.2 Inhaltsinformation (Content Information)

7.2.1 Anforderung

Die präzise Definition der zu erhaltenden Inhaltsinformation ist essentiell für ein Archiv, da nur so ermittelt werden kann ob eine Erhaltungsmassnahme alle oder nur Teile der zu erhaltenden Inhaltsinformationen bewahrt. Diese Definition hat entsprechend Auswirkungen auf die weitere Definition von Arbeitsabläufen, AIP, Repräsentationsinformation etc.

Die Anforderungen gemäss CCSDS 652.0-M-1 Audit sind:

> *4.1.1 The repository shall identify the Content Information and the Information Properties that the repository will preserve.*
>
> *Supporting Text*
> *This is necessary in order to make it clear to funders, depositors, and users what responsibilities the repository is taking on and what aspects are excluded. It is also a necessary step in defining the information which is needed from the information producers or depositors.*
>
> *Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement*
> *Mission statement; submission agreements/deposit agreements/deeds of gift; workflow and Preservation Policy documents, including written definition of properties as agreed in the deposit agreement/deed of gift; written processing procedures; documentation of properties to be preserved. [CCSDS652] Seite 4-1*

7.2.2 Definition

Das Public Domain Projekt übernimmt die Verantwortung die ihr übertragenen digitalen Audiowerke zu erhalten. Die Inhaltsinformation von Audiowerken wird definiert als:

- Die akustische Information im Frequenzbereich der von Menschen hörbar ist (15 Hz bis 20 kHz)
- Alle nötigen Begleitinformationen (Metadaten) um die Identität, die Herkunft, die Entstehung und die Authentizität zu bestimmen.

8 Anforderungsanalyse

Dieses Kapitel beschäftigt sich mit den Anforderungen an die konkrete Ausgestaltung des Langzeitarchivs für digitale Audiowerke. Es bildet die Anwendung des in Kapitel 4 vorgestellten Referenzmodells für das Public Domain Projekt. Die Priorisierung der Umsetzung erfolgt mit Hilfe der Resultate des Audits aus Kapitel 6.

Wie im Kapitel 4.4 erläutert wurde, ist Langzeiterhaltung ein kontinuierlicher Prozess, in dem regelmässig das technische, soziale und politische Umfeld des Archivs beobachtet werden muss und bei Bedarf auf veränderte Bedingungen oder Anforderungen reagiert werden muss.

Alle Entscheidungen sind so zu gestalten, dass die Langzeiterhaltung nicht gefährdet wird.

8.1 Repräsentationsinformation

Wie aus dem Audit hervorgeht, war dem Public Domain Projekt bisher nicht bewusst, was unter Repräsentationsinformation zu verstehen ist und deren Wichtigkeit für die Langzeiterhaltung. Entsprechend kommt das Projekt der Anforderung nicht nach, die nötige Repräsentationsinformation aktiv zu bestimmen und die Verantwortung für deren Erhalt zu übernehmen. Trotz des konsequenten Einsatzes von Open Source Software im Public Domain Projekt, was die Risiken zu grossen Teilen minimiert, kann dieses Thema nicht vernachlässigt werden. Wie schon im OAIS Referenzmodell erkannt wurde, kann die nötige Repräsentationsinformation sehr schnell ein grösseres Netzwerk von Informationen umfassen, die nötig sind, um den Inhalt einer Datei in eine aussagekräftige Form zu bringen.

Die Anforderung ist, dass der Detailgrad des vorhandenen und abrufbaren Wissens bis auf das letzte Bit herunter reichen muss. Das beinhaltet das Dateiformat und dass jeder Bestandteil der Datei (Metadaten, Prüfsummen, eingebettete Bilder, Zeichencodierung, Bit und Byte Reihenfolge, Filesystem, Zugriffsprogramme etc.) bekannt sein muss.

8.2 Strategie zur Bewältigung des Technologiewandels

Im Kontext von digitalen Audiowerken hat die Transformation (Siehe Kapitel 4.3.1) Vorteile. Es existieren heutzutage mehrere Dateiformate die sich als Basis für ein AIP für Audiowerke eignen und die eigentliche Audioinformation bleibt dabei die selbe (lineare Pulse Code Modulation, PCM), unterschiedlich ist die Unterstützung von eingebetteten Metadaten. Die Transformation ermöglicht auch einen einfacheren Zugang der vorgesehenen Zielgruppe zum AIP, denn es wird keine aufwändige Emulationsumgebung benötigt um ein Auslieferungsinformationspaket (DIP) zu erstellen. In anderen Bereichen z. B. Langzeiterhaltung von Computerprogrammen (inkl. Computerspiele und computerbasierte Kunst) ist der Einsatz von Emulation ein aktives Forschungsthema.

Gefordert wird, dass Definitionen und Systeme so vorausgeplant werden, dass eine Transformation zu jedem gegebenen Zeitpunkt möglich ist.

8.3 Nachweis der Authentizität

Wie beschrieben muss die vorgesehene Zielgruppe sich von der Authentizität eines digitalen Objektes überzeugen können. Zu diesem Zweck muss das Public Domain Projekt dazu benötigte Erhaltungsmetadaten bereitstellen.

Die vorgesehene Zielgruppe hat folgende Anforderungen:

- Die Herkunft des Werks und Beteiligten sind einsehbar und mit Quellen belegt
- Zur eindeutigen Unterscheidung sind Referenzen zu Normdateien (authority files) vorhanden
- Abrufbar ist die Geschichte und alle Änderungen eines digitalen Objektes, seit es in der Obhut des Public Domain Projektes ist
- Es ist ein Nachweis verfügbar, dass die Daten nicht auf ungewollte Weise verändert wurden

8.4 Anforderung an Metadaten

Die Anforderungen die vom OAIS Referenzmodell und vom Audit an die Metadaten gestellt werden sind weitreichend und detailliert und wurden in Kapitel 4.5 *Klassen von Metadaten* beschrieben. Diese Anforderungen müssen von den Metadatenstandards im Public Domain Projekt erfüllt werden

Zusätzlich werden vom Public Domain Projekt die folgenden Anforderungen an die Metadaten gestellt:

- Tauglich für die Beschreibung von Audiowerken
- Standardisierte und verbreitete Metadaten zur Erhöhung der Interoperabilität (Datenaustausch, Software Unterstützung) und Vereinfachung der Nutzung durch die vorgesehene Zielgruppe
- Die Metadaten sollen maschinenlesbar und -interpretierbar sein

8.5 Anforderung an die Definition des Archivinformationspakets (AIP)

Die Definition des Archivinformationspakets (Archival Information Package, AIP) ist eine anspruchsvolle Aufgabe, da an ein AIP viele Anforderungen gestellt werden, die alle abgedeckt werden müssen. Dazu kommt, dass andere Definitionen, wie die zu erhaltende Inhaltsinformation oder die verwendeten Metadatenstandards, einen grossen Einfluss auf die AIP Definition haben.

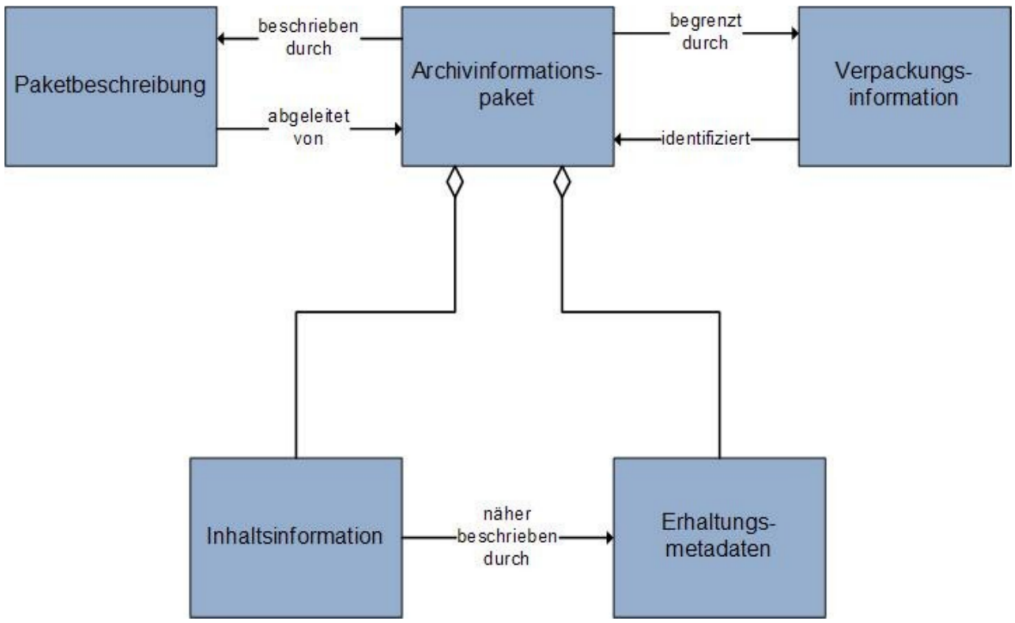

Abbildung 19: Modell des Archivinformationspakets (AIP) [NES13], Seite 66

Die AIP Definition wird langfristig nicht die selbe bleiben, sondern sich anpassen müssen. Wenn es die Umstände nötig machen oder weil sich die Prozesse in einem Archiv dadurch vereinfachen lassen, kann ein Nachfolge-AIP definiert werden (das wiederum alle AIP Anforderungen erfüllt) und der bestehende Archivbestand per Transformation migriert werden.

Im OAIS Referenzmodell wird ein AIP definiert als:

> *Das AIP ist definiert als eine präzise Art, auf einen Satz von Informationen zu verweisen, der im Prinzip alle notwendigen Eigenschaften für eine permanente, oder unbegrenzte, Langzeiterhaltung eines bestimmten Informationsobjekts hat. Das AIP ist selbst ein Informationsobjekt, das ein Container für andere Informationsobjekte ist. Innerhalb des AIPs befindet sich das bestimmte Informationsobjekt, das hier "Inhaltsinformation" genannt wird. Innerhalb des AIPs befindet sich außerdem ein Informationsobjekt, das "Erhaltungsmetadaten" genannt wird. Diese Erhaltungsmetadaten enthalten zusätzliche Information über die Inhaltsinformation und werden gebraucht, um die Inhaltsinformation für eine unbestimmt langfristige Zeit aussagekräftig zu machen. [NES13] Seite 66*

Die IASA stellt weitere Anforderungen an die Auswahl der Dateiformate:

> *Digitale Kodierungsschemata (Formate) sowie digitale Auflösungen sind Gegenstand weiterer Entwicklungen. Unabhängig davon sei betont, dass für Archivierungszwecke nur offen definierte Formate verwendet werden dürfen und keinesfalls proprietäre, die nur von wenigen Herstellern unterstützt werden. Echte Fileformate (Datenformate) sind digitalen Audiostreams (R-DAT, CD-Audio) hinsichtlich der Datensicherheit und deren Überwachung überlegen. [TC03] Kapitel 10*

Die grundlegenden Anforderung an ein AIP sind somit:

- Die definierte Inhaltsinformation kann vollständig vom AIP aufgenommen werden
- Aus den Informationen im AIP muss bestimmt werden können, welche Repräsentationsinformation benötigt wird um die im AIP enthaltenen Datenobjekte in eine aussagekräftige Form zu bringen
- Die Anforderungen der Abschnitte 8.3 *Nachweis der Authentizität* und 8.4 *Anforderung an Metadaten* müssen erfüllt werden
- Die Dateiformate sind offen spezifiziert

Seit der Definition der ersten Version des OAIS Referenzmodells sind aus der praktischen Anwendung zusätzliche Anforderungen an ein AIP diskutiert worden. Für das Public Domain Projekt wird zusätzlich gefordert:

- Die Möglichkeit das Datenobjekt unabhängig von den Erhaltungsmetadaten auf ungewollte Veränderungen zu überprüfen. Das Datenobjekt wird nach der AIP Erstellung nicht mehr verändert, die Metadaten werden aber im Laufe der Zeit ergänzt. Eine unabhängige Überprüfung erleichtert somit den Nachweis der Authentizität des Datenobjekts.
- Das AIP soll eine einzelne Datei sein, damit das zur Verfügung Stellen im Internet vereinfacht wird und die Metadaten immer mit dieser Datei weitergegeben werden. Diese Anforderung ist relativ neu und kommt mehr aus den Endnutzeranforderungen als aus den Archivanforderungen: Sobald die aus Endnutzersicht spannende Musik und die langweiligen Metadaten separate Dateien sind, werden bei der ersten Gelegenheit die Metadatendateien gelöscht oder nicht weitergegeben, was dann (viel später) zu Problemen führen kann, die Herkunft und Authentizität eines Werks zu bestimmen.

8.6 Archivspeicher (Archival Storage)

Der Archivspeicher ist eine der Kernkomponenten in der Langzeiterhaltung, dessen Anforderungen vor allem durch das Risikomanagement der Datenspeicherung auf unzuverlässigen Medien und durch die Sicherstellung der Interpretierbarkeit der Daten vorgegeben sind.

Im Audit gibt es starke Argumente um freier Software gegenüber proprietärer den Vorzug zu geben:

> *5.1 TECHNICAL INFRASTRUCTURE RISK MANAGEMENT*
>
> *5.1.1 The repository shall identify and manage the risks to its preservation operations and goals associated with system infrastructure.*
>
> *Supporting Text*
> *This is necessary to ensure a secure and trustworthy infrastructure.*
>
> *Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement*
> *[...]use of strongly community supported software e.g., Apache, iRODS, Fedora[...]*
>
> *Discussion*
> *[...]The repository should provide mechanisms that minimize risk from dependencies on proprietary or obsolete system infrastructure and from operational error. The degree of support required relates to the mechanisms that minimize risk from dependencies on proprietary or obsolete system infrastructure and from operational error. The degree of support required relates to the criticality of the subsystem(s) involved in long-term preservation.[...]*
> *[CCSDS652] Seite 5-1*

Die Anforderungen an den Archivspeicher für das Public Domain Projekt sind:

- Verhindern von unerwünschten Veränderungen
- Die benötigte Repräsentationsinformation bis hinunter zum Bit-Level muss vorhanden sein
- Es muss nachvollziehbar sein wer, was, wann und wieso hinzugefügt oder geändert hat → Versionsverwaltung der gespeicherten Daten
- Risikomanagement für alle denkbaren Vorkommnisse, die die Daten gefährden könnten, dazu gehört mindestens ein Backup an einem geographisch entfernten Ort
- Einsatz von Free and Open Source Software (FOSS) um die Risiken der Abhängigkeit von proprietären Systemen zu minimieren

8.7 Übernahmeprozess (Ingest)

Das Ziel des Übernahmeprozesses ist das Entgegennehmen von neuem digitalen Archivgut und allen nötigen Schritten um daraus ein AIP zu generieren. Dazu gehört vor allem das Prüfen und Vervollständigen der Erhaltungsmetadaten. Dies dauert so lange bis alle zwingend erforderlichen Metadaten für ein AIP vorhanden sind. Danach wird per definiertem Prozess ein AIP generiert und dem Archivspeicher übergeben.

> *4.1.5 The repository shall have an ingest process which verifies each SIP for completeness and correctness. [CCSDS652] Seite 4-4*

Je nach Produzent sind unterschiedliche Übernahmeprozesse zu definieren. Derzeit liegt der Fokus auf dem Digitalisierungsprozess der Freiwilligen im Public Domain Projekt. Andere denkbare Produzenten könnten z. B. andere Archive oder Produktionsstudios sein.

Im aktuellen Prozess im Public Domain Projekt gibt es keinen Unterschied zwischen einem SIP und dem finalen AIP. Das führt dazu, dass Endnutzer nicht feststellen können, ob es sich um ein finales AIP handelt oder ob es sich noch um ein unfertiges SIP handelt. Dies ist offensichtlich problematisch und muss behandelt werden.

Die Anforderung daraus ist, dass unterscheidbar sein muss, ob es sich um ein SIP (Aufnahmeprozess läuft noch) oder um ein AIP (Validiertes, vollständiges Informationsobjekt) handelt. Aussenstehende Nutzer interessieren sich im Allgemeinen für finale AIPs und die daraus erstellten DIPs (Altagstaugliches Dateiformat).

Die Freiwilligen im Projekt wollen wissen, welche SIPs vorhanden sind und in welchem Zustand: z. B. welche Informationen fehlen einem SIP noch und müssen noch recherchiert werden bevor es ins Archiv transferiert werden kann. Beim Public Domain Projekt, wo das Archivgut aus dem Digitalisierungsprozess stammt, sind im Übernahmeprozess viele manuelle arbeiten wie Literaturrecherche nötig. Der Übernahmeprozess kann aber von automatisierten Hilfsmitteln unterstützt werden z. B. um SIPs zu markieren die noch unvollständig sind.

Ein solches Hilfsmittel soll auch prüfen, ob ein SIP den Empfehlungen der IASA für Samplingrate und Ampltudenauflösung entspricht:

> *Die Auflösung digitaler Formate wird begrenzt durch die definierte und endliche Abtastrate sowie die digitale Wortlänge. Während für digital aufgenommene Signale die Originalauflösung im digitalen Archivformat beibehalten werden soll, stellt die Auflösung für analoge Signale immer einen Kompromiss dar. Prinzipiell sind hohe digitale Auflösungen für eine adäquate Wiedergabe aller subtilen Details der analogen Originalsignale zu wählen. [...] Zurzeit sind Analog-Digital-Konverter mit einer Abtastrate von 192 kHz und einer Amplitudenauflösung von 24 bit Standard. Für analoge Originale empfiehlt IASA als digitale Mindestauflösung 48 kHz Abtastrate und 24 bit Wortlänge. In Gedächtnisinstitutionen wird weitgehend eine Auflösung von 96 kHz/24 bit angewendet. Bessere Übertragungen der unerwünschten Teile eines Tondokuments machen die spätere Entfernung dieser Artefakte durch digitale Signalverarbeitung bei der Herstellung von Benützerkopien leichter. Wegen des transienten Charakters der Konsonanten sind Sprachaufnahmen Musikaufnahmen gleichzustellen. [TC03] Kapitel 10*

8.8 Anforderungen der vorgesehenen Zielgruppen

Nach der Definition der vorgesehenen Zielgruppen ist der nächste Schritt herauszufinden, welche Erwartungen diese Zielgruppen haben und wie diese das Public Domain Projekt nutzen möchten.

Um herauszufinden ob sich diese Anforderungen wirklich mit den Erwartungen der Zielgruppen decken, können in den folgenden Jahren Umfragen durchgeführt werden.

Die Zielgruppe *Allgemeinheit* (Global Community) möchte schnell herausfinden was es im Archiv gibt und möchte die Musik auch gleich anhören und herunterladen können. Angeboten werden sollen Formate die aktuell übliche PC und Smartphone Betriebssysteme ohne Zusatzsoftware abspielen können.

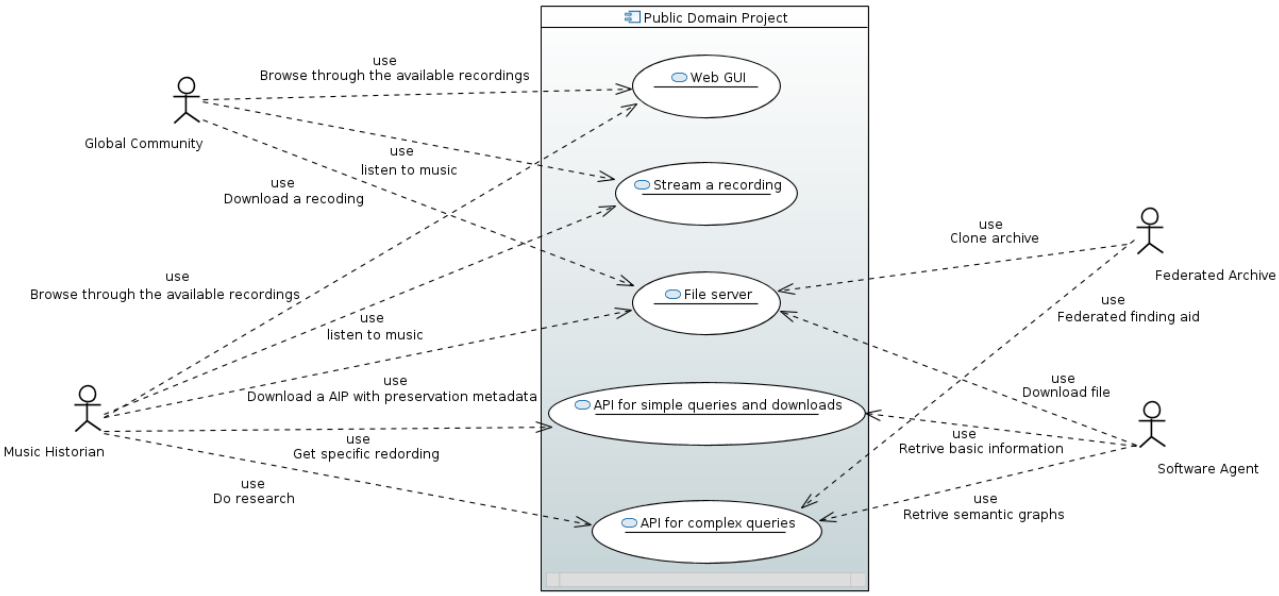

Abbildung 20: UseCase Diagramm der vorgesehenen Zielgruppen

Die Zielgruppe *Musikwissenschaftler, Historiker und Interpretationsforscher* hat zuerst mal die selben Anforderungen wie die Allgemeinheit aber zusätzlich die Anforderung an detaillierte Suchmöglichkeiten. So soll gezielt in einer Auswahl von Genres, Zeitabschnitten, Urhebern, Interpreten, geographischen Orten etc. gesucht werden können. Es sollen Listen mit allen verfügbaren Werken von einem Urheber oder Interpreten etc. zugänglich sein. Es muss möglich sein, die Daten aus dem Public Domain Projekt mit Daten aus anderen Quellen zu verknüpfen. Dazu ist es nötig, dass die Daten Identifikatoren aus anerkannten Normdateien zur eindeutigen Identifizierung enthalten. Es soll möglich sein, auf die vollständigen AIPs zugreifen zu können.

Die Zielgruppe *Suchmaschinen, Metaarchive, Datenanalyseprogramme (Bots)* , besteht nicht aus Menschen sondern aus Maschinen und hat entsprechend andere Anforderungen an die Durchsuchbarkeit und Darstellung der Daten. Zu dieser Zielgruppe gehören auch die Abfragesysteme der mit dem Public Domain Projekt verbundenen Archive. Es muss eine entsprechende Softwareschnittstelle (Application programming interface, API) für einfache Suchanfragen bestehen mit dessen Antworten dann Audiowerke heruntergeladen oder die Erhaltungsmetadaten abgerufen werden können. Des weiteren besteht der Wunsch nach einer Möglichkeit komplexe Anfragen zu stellen (Z. B. Liste aller Werke von 1730 bis 1750 von Schülern von Johann Sebastian Bach bei der eine Orgel unter den Instrumenten ist) und dessen Antworten Referenzen auf maschineninterpretierbare Konzepte und Normdaten beinhaltet (Semantic Web, Linked Data). Der aktuelle Stand der Technik für solche Anfragen ist SPARQL[20], dabei werden die Daten im Format RDF 1.1, serialisiert als RDF/XML, zurückgeliefert.

[20] https://www.w3.org/TR/sparql11-query/

9 Metadatenstandards

In diesem Kapitel werden die Ergebnisse der Recherchen zu Metadatenstandards präsentiert. Der Fokus wird nur auf die Metadaten gelegt und nicht auf die zu Grunde liegenden Basistechnologien wie XML und RDF. Für das Verständnis dieses Kapitels ist ein rudimentäres Basiswissen zu Semantic Web und RDF hilfreich. Es wurde davon abgesehen dieses Wissen in diesem Bericht zu vermitteln, da es den Rahmen dieses Berichts sprengen würde und es ausserhalb des Themas Langzeiterhaltung liegt.

Einen Überblick vermitteln folgende Webseiten:

- Basics of the Semantic Web http://strangelove.netlabs.org/semantic-web-basics/
- RDF 1.1 Primer https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

9.1 Aktueller Stand der Metadaten im Public Domain Projekt

Das Public Domain Projekt erfasst ziemlich detaillierte Metadaten, diese sollten aber noch maschinenlesbar gemacht werden und mit Feldern aus Metadatenstandards abgeglichen werden. Die derzeit vom Public Domain Projekt erfassten Metadaten sind als Liste im Anhang wiedergegeben.

Die ursprüngliche Idee, die Erhaltungsmetadaten vollständig in die Flac Datei einzubetten, musste aufgegeben werden, da dies sehr aufwändige Erweiterungen in der Metadatenspezifikation von Flac erfordert hätte. Entsprechend müsste dann auch alle Software zur AIP Generierung und Änderung etc. selber implementiert werden, weil das in Flac bis zu diesem Zeitpunkt nicht vorgesehen ist. Im Anhang die kurze aber vollständige Liste der offiziellen Metadatenfelder in der Vorbis Spezifikation, welche auch für Flac Dateien gilt.

9.2 Mögliche Standards um fehlende Erhaltungsmetadaten zu erfassen

Es wurde viel Zeit in die Suche und den Vergleich von Metadatenstandards investiert, um herauszufinden was andere Archive, Bibliotheken oder andere Branchen zur Langzeiterhaltung verwenden.

Es hat sich gezeigt, dass derzeit sehr viele solcher Standards und Vokabulare existieren. Diese sind dann vielfach entweder sehr eingeschränkt, nicht maschineninterpretierbar oder spezifisch für einen Fachbereich. Viele Vokabulare sind regional oder auf bestimmte Sprachen/Fachbereiche limitiert. Im Bereich der Archive der Audiovisuellen Medien sind auch verschiedene Standards in Verwendung ohne das sich eine Konsolidierung abzeichnet.

Es hat sich gezeigt, dass es nicht ausreichen wird, für alle Erhaltungsmetadaten einen einzelnen Metadatenstandard zu nutzen, sondern dass mehrere solche Standards kombiniert werden müssen um die Anforderungen zu erfüllen.

9.2.1 Nicht weiterverfolgte Metadatenstandards

Im Verlauf der Projektarbeit wurden unter anderem folgende Standards evaluiert und danach nicht mehr weiterverfolgt:

- Library of Congress Subject Headings (LCSH), nicht maschinenlesbar
- Mets, *[...] providing an encoding format for descriptive, administrative, and structural metadata for textual and image-based works.* *http://www.loc.gov/standards/mets/METSOverview.v2.html*
- MARC21, MAchine Readable Cataloguing, nicht allgemein menschenlesbar da die Kategorisierung nur aus Nummern zusammengesetzt wird
- foaf, ist in PREMIS inkludiert
- Pbcore, kleine Verbreitung, nicht kompatibel mit RDF/LinkedData
- bbcore, Verwendung innerhalb der Britisch Broadcast Corporation (BBC)
- EBUCore, verwendet von Memoriav, mehr orientiert an Produktionsabläufen von Radio und Fernsehstudios als Langzeiterhaltung, darum zu überladen für das Public Domain Projekt
- SKOS, ist in PREMIS inkludiert
- Functional Requirements for Bibliographic Records (FRBR), fokusiert auf Bücher in Bibliothekskatalogen, foaf ist zu einem Teil äquivalent mit den FRBR Metadaten
- ISAD(G), *[...] does not itself provide a machine-readable binding* *http://www.ukoln.ac.uk/metadata/dcmi/collection-provenance/*, fokusiert auf Aktenarchive
- Data Catalog Vocabulary (DCAT), Austauschformat für Datenverzeichnisse, nicht passend
- XMP, nur für Bild- und PDF-Dateien geeignet
- IPTC, nur für Bilddateien geeignet
- Ontology for Media Resources (ma-ont), wird nur von EBUcore referenziert, kaum verbreitet

9.2.2 DublinCore, DC

Das Kernelement Set *DublinCore* ist aus dem Bedürfnis entstanden, Katalogbestände über verschiedene Institutionen auszutauschen, was nicht einfach so geht, weil jede Institution anderen Metadatenvorgaben folgt. Das Core im Namen bezieht sich darauf, dass es sich hier um ein Set von 15 Basisfeldern geht die einem minimalen Konsens entsprechen, welche Begleitinformation bei jedem Werk vorhanden ist. Es bildet daher die Basis für heutige Austauschformate. Entsprechend ist klar, dass DublinCore nicht gross genug ist, um alle Erhaltungsmetadaten aufzunehmen. Damit können Herkunftsinformationen (Provenance Information) und ein Teil der Kontextinformation (Context Information) abgedeckt werden. Es besteht zudem das Problem, dass zwar die 15 Felder definiert sind, aber nicht strikt das Format oder Vokabular des Feldinhalts.

In der Zwischenzeit wurde versucht die (bewusst) minimal gehaltene Spezifikation zu erweitern und die Popularität von DublinCore für weitere Standardisierung und Interoperabilität zu nutzen:

> *Early Dublin Core workshops popularized the idea of "core metadata" for simple and generic resource descriptions. The fifteen-element "Dublin Core" achieved wide dissemination as part of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and has been ratified as IETF RFC 5013, ANSI/NISO Standard Z39.85-2007, and ISO Standard 15836:2009.*
>
> *Starting in 2000, the Dublin Core community focused on "application profiles" -- the idea that metadata records would use Dublin Core together with other specialized vocabularies to meet particular implementation requirements. During that time, the World Wide Web Consortium's work on a generic data model for metadata, the Resource Description Framework (RDF), was maturing. As part of an extended set of DCMI Metadata Terms, Dublin Core became one of most popular vocabularies for use with RDF, more recently in the context of the Linked Data movement.*
>
> *The consolidation of RDF motivated an effort to translate the mixed-vocabulary metadata style of the Dublin Core community into an RDF-compatible DCMI Abstract Model (2005). The DCMI Abstract Model was designed to bridge the modern paradigm of unbounded, linked data graphs with the more familiar paradigm of validatable metadata records like those used in OAI-PMH. A draft Description Set Profile specification defines a language for expressing constraints in a generic, application-independent way. The Singapore Framework for Dublin Core Application Profiles defines a set of descriptive components useful for documenting an application profile for maximum reusability. Quelle: http://dublincore.org/metadata-basics/*

Trotz der längeren Zeit seit Beginn dieser Arbeiten und der weiter gestiegenen Popularität von DublinCore existieren nur für wenige Fachbereiche definierte Applikationsprofile. Es existiert derzeit kein Applikationsprofil für Audiowerke. Der Metadatenstandard EBUCore[21] veröffentlicht von der European Broadcasting Union (EBU) ähnelt einem Applikationsprofil, definiert aber einen eigenen Namensraum, was der Interoperabilität abträglich ist. Er wurde 2008 erstmals publiziert und hat bisher nur eine kleine Verbreitung gefunden.

[21] https://tech.ebu.ch/MetadataEbuCore

9.2.3 PREMIS 3.0

Bei PREMIS (Preservation Metadata: Implementation Strategies) handelt es sich um einen Standard der gezielt für Erhaltungsmetadaten entwickelt wurde:

> *The PREMIS Data Dictionary and its supporting documentation is a comprehensive, practical resource for implementing preservation metadata in digital archiving systems. The Data Dictionary is built on a data model that defines five entities: Intellectual Entities, Objects, Events, Rights, and Agents.*[22]

Der Standard ist bei der US Amerikanischen Library of Congress angesiedelt und wird von der internationalen PREMIS-Arbeitsgruppe vorangetrieben, die ursprünglich vom Online Computer Library Center (OCLC) und der Research Libraries Group (RLG) ins Leben gerufen wurde. Die erste Version wurde 2005 präsentiert und ist seitdem aktiver Gegenstand von Forschung und Entwicklung im Archivbereich. Die Version 3.0 wurde 2015 präsentiert und erweitert die Möglichkeiten Abhängigkeiten von Repräsentationsinformationen zu repräsentieren und gibt detailliertere Informationen zu Agenten.

In der überarbeiteten Version des OAIS Referenzmodells von 2012 wird PREMIS unter den Referenzen aufgeführt. [NES13] Seite 5

PREMIS baut stark auf schon bekannte und akzeptierte Ontologien auf, was die Integration in bestehende SemanticWeb Umgebungen vereinfachen sollte und die Interoperabilität von verschiedenen Quellen wesentlich fördert. Abbildung 21 zeigt die Verwandschaftsbeziehung von PREMIS zu anderen Ontologien. Wie zu sehen ist, wird vom DublinCore Vokabular dcterms geerbt.

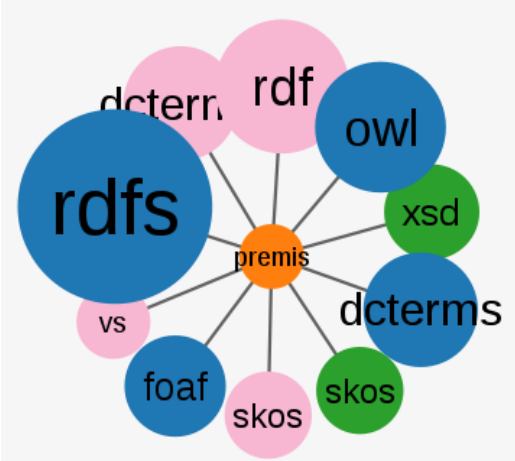

Abbildung 21: Übersicht der Verwandschaft von PREMIS mit anderen Ontologien. Quelle: http://lov.okfn.org/dataset/lov/vocabs/premis

Legende zur Abbildung 21:

- Friend of a Friend vocabulary (foaf) http://www.foaf-project.org/
- Simple Knowledge Organization System (skos) http://www.w3.org/2009/08/skos-reference/skos.html
- SemWeb Vocab Status ontology (vs), an RDF vocabulary for relating SemanticWeb vocabulary terms to their status. http://www.w3.org/2003/06/sw-vocab-status/ns

[22] http://www.loc.gov/standards/premis/v3/index.html

Abbildung 22 zeigt ein Beispiel, wie ein Audiowerk im Besitz des Public Domain Projekts in PREMIS aussehen könnte.

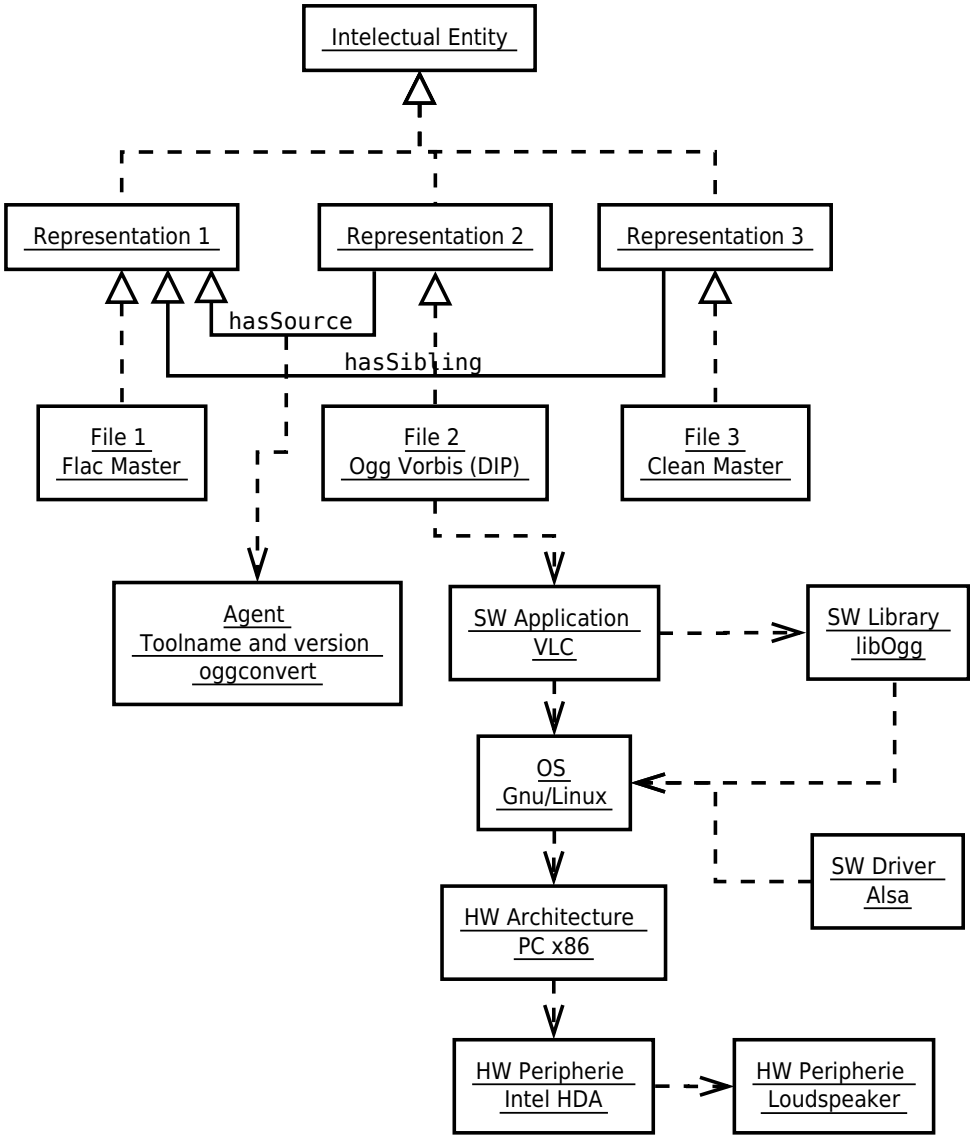

Abbildung 22: Beispiel wie im Public Domain Projekt mit PREMIS 3.0 die Provenienzinformationen und die Abhängigkeiten zu Repräsentationsinformationen modelliert werden kann.

Für den praktischen Einsatz müssen auch bei PREMIS die verwendeten Vokabulare festgelegt werden, da in der Spezifikation nur wenige Vokabulare von PREMIS als Referenzen angegeben werden und diese nicht als zwingend vorgeschrieben werden. Im Public Domain Projekt werden diese Referenzen als starke Empfehlungen interpretiert und entsprechend versucht, diese bei der Implementierung von PREMIS zu verwenden.

9.3 Eindeutige Identifikatoren, Normdatei (Authority control)

Es wurde schon mehrfach auf eindeutige Identifikatoren verwiesen, die nötig, sind um einzelne Werke und deren Repräsentationen eindeutig zu identifizieren und voneinander zu unterscheiden. Sie dienen auch dazu, Personen wie Autoren und Interpreten, unverwechselbar zu machen. Die Erstellung und Katalogisierung solcher Identifikatoren ist ein eigener Aufgaben- und Forschungsbereich des Bibliotheks- und Archivwesens. Der deutschsprachige Begriff für ein Identifikationssystem und der zugehörige Katalog von bekannten Identifikatoren ist *Normdatei*, das englischsprachige Äquivalent heisst *authority control*. Durch die Verwendung von Identifikatoren aus externen, akzeptierten Normdateien in den Erhaltungsmetadaten kann die vorgesehene Zielgruppe die Identität und die Authentizität eines digitalen Objekts besser und zuverlässiger Nachvollziehen.

Aus dem Audit geht hervor, dass das aktuell verwendete Namensschema vorhersehbare Probleme hat, die derzeit noch nicht zum Tragen kommen. Hauptsächlich sobald andere Medien als Schellackplatten aufgenommen werden sollen. Es ist entsprechend angebracht, Gedanken darüber anzustellen, ob ein alternatives Namensschema gleich eindeutige Identifikatoren aus einer Normdatei übernehmen oder integrieren soll. Es stellt sich dabei natürlich schnell die Frage, welche Normdateien für das Public Domain Projekt taugliche Kandidaten wären.

Während den Untersuchungen wurden folgende Normdateien verworfen: DNB, Discogs, WikiData und Label Code (Bei Tonträgern ähnlich zu ISBN).

9.3.1 VIAF

VIAF ist die internationale Referenz, um Personen eindeutig zu Identifizieren und ist entsprechend aktuell zu bevorzugen:

> *The VIAF® (Virtual International Authority File) combines multiple name authority files into a single OCLC-hosted name authority service. The goal of the service is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web. Quelle: https://viaf.org/*

9.3.2 MusicBrainz

> *MusicBrainz is an open music encyclopedia that collects music metadata and makes it available to the public.vMusicBrainz aims to be:*
>
> 1. *The ultimate source of music information by allowing anyone to contribute and releasing the data under open licenses.*
>
> 2. *The universal lingua franca for music by providing a reliable and unambiguous form of music identification, enabling both people and machines to have meaningful conversations about music.*
>
> *Like Wikipedia, MusicBrainz is maintained by a global community of users and we want everyone — including you — to participate and contribute. Quelle: https://musicbrainz.org/*

MusicBrainz gehört derzeit zu den grössten Datenbanken im Bereich Musik, fokussiert auf Tonträger und allen Beteiligten, also Autoren, Interpreten, Verlage. Das Projekt basiert auf modernen Ansätzen zu Metadatenstandards wie der Trennung von intellektuellem Werk und deren Repräsentationen und

das Nutzen eines formal definierten verlinkten Graphen als Datenmodell. Die Datenbank ist unter Creative Commons Zero Restrictions (CC0) lizenziert.

MusicBrainz legt viel Wert auf Qualität, entsprechend existieren Freigabeprozesse für neue Einträge und Änderungen. Als weiteres Qualitätsmerkmal werden, wenn möglich und vorhanden, Referenzen zu anderen Normdateien wie VIAF und WikiData integriert.

Ein Problem für das Public Domain Projekt ist, dass die Abdeckung für alte Tonträger wie Schellackplatten schlecht ist, weil kaum Freiwillige in MusicBrainz aktiv sind, die in deren Besitz sind. Die Abklärungen während der Projektarbeit haben ergeben, dass MusicBrainz grundsätzlich sehr offen ist für eine Partnerschaft, da sie an den Metadaten der alten Tonträgern des Public Domain Projekts sehr interessiert sind um die eigene Abdeckung zu erhöhen.

Es sind weitere Abklärungen notwendig, um herauszufinden wie eine solche Zusammenarbeit aussehen könnte und wie ein dazu passender Arbeitsablauf aussehen würde.

9.4 Was von der Zielgruppe erwartet wird

Wenn schon kein Konsens unter den Archiven herrscht, dann lässt sich vielleicht eine Antwort finden, wenn man von der anderen Seite her kommt. Mit welchen Metadatenstandards und Vokabularen können die vorgesehenen Zielgruppen des Public Domain Projekts umgehen und welchem würden sie den Vorzug geben?

Eine Antwort ist sicher DublinCore mit dem minimalen Set von 15 Feldern. DublinCore stellt sicherlich den gemeinsamen Nenner dar, der der Zielgruppe schon viele Möglichkeiten eröffnet um Daten aus mehreren Quellen abzufragen. Damit diese verknüpft werden können, müssen eindeutige Identifikatoren enthalten sein. Je nach Quelle werden andere Normdateien verwendet aber mit Aufwand (Abfragen der Normdateien und abgleichen der Daten mit diesen Resultaten) lassen sich Daten von verschiedenen Quellen trotzdem vereinen.

Egal welche Standards am Ende eingesetzt werden, deren Verwendung muss genau dokumentiert werden, damit die Metadaten konsistent erfasst werden und damit die Zielgruppen nachlesen können, wie sie per API oder SPARQL gezielt detaillierte Suchanfragen stellen können.

10 Empfehlungen und weiterführende Arbeiten

In diesem Kapitel geht es darum, mit den Informationen aus allen vorhergegangenen Kapiteln konkrete Vorschläge für das Public Domain Projekt und die Schweizerische Stiftung Public Domain auszuarbeiten.

10.1 Allgemeine Empfehlungen

Die neu erstellten Definitionen sollen, sobald sie vom Stiftungsrat und den Projektmitgliedern abgesegnet worden sind, im MediaWiki eingepflegt werden und in die anderen Sprachversionen übersetzt werden.

Sobald innerhalb des Public Domain Projekts Software entwickelt wird, muss ein System zur Verwaltung der Softwareentwicklung (Versionierung, Dokumentation etc.) evaluiert und eingesetzt werden.

10.2 Vorschlag für das Archivinformationspaket (Archival Information Package, AIP)

Wie im Audit festgestellt wurde, ist das derzeit verwendete AIP, eine Flac Datei zusammen mit einer Beschreibungsseite im MediaWiki, zu wenig tauglich um die Anforderungen die an ein AIP gestellt werden zu erfüllen. Im jetzigen AIP werden die Erhaltungsmetadaten nur als menschenlesbarer Text innerhalb einer Datenbank gespeichert, was es erschwert die nötige Repräsentationsinformation vorzuhalten. Ein weiteres Problem des jetzigen AIP ist die räumliche Trennung der Flac Datei und der Datenbank, diese werden auf zwei verschiedenen Serversystemen gespeichert.

Zu den sehr gängigen Methoden gehört, dass zu einem AIP mehrere Dateien gehören: Das eigentliche Datenobjekt und in separaten Dateien die Erhaltungsmetadaten. Das AIP wird dann definiert als Ordnerstruktur oder als Zip-Archiv mit eindeutiger Kennung im Namen. Für das Public Domain Projekt, das sich mit audiovisuellem Kulturgut auseinandersetzt, wird ein anderer Vorschlag gemacht.

Wie gefordert, soll beim AIP für das Public Domain Projekt verhindert werden, dass beim einfachen Herunterladen und Kopieren eines AIP Metadaten verloren gehen. Der Vorschlag ist eine Weiterentwicklung der ursprünglichen Idee hinter der Verwendung von Flac Dateien (Offene Spezifikation, direkt abspielbar) um die Limiten der jetzigen Lösung auszumerzen.

Es wird vorgeschlagen, eine AIP Definition zu entwickeln auf Basis des Matroska Container Formats:

> *Matroska the extensible, open source, open standard Multimedia container. Matroska is usually found as .MKV files (matroska video), .MKA files (matroska audio) and .MKS files (subtitles) and .MK3D files (stereoscopic/3D video). It is also the basis for .webm (WebM) files.*[23]

Matroska Container sind für das Public Domain Projekt derzeit ein interessanter Ansatz um eine über längere Zeit nutzbare AIP Definition mit Erweiterungsmöglichkeiten in der Zukunft zu haben. Begründet wird dies durch folgende Eigenschaften:

- Im Gegensatz zu Zip-Archiven sind sie in einem Mediaplayer der Matroska und Flac unterstützt direkt abspielbar (Verfügbar für alle aktuell gängigen Betriebssysteme).
- Die Matroska Container Spezifikation unterstützt Flac als Audiocodec seit mindestens 2010[24]

[23] https://matroska.org/
[24] https://matroska.org/node/16/revisions/16/view

- Matroska Container haben die Möglichkeit beliebige Dateien einzubetten. Somit können XML Dateien mit den in Kapitel 9.2 vorgeschlagenen DublinCore und PREMIS Metadaten hinterlegt werden. Weitere Informationen wie z. B. die gespielten Noten als MIDI Dateien könnten so auch hinzugefügt werden. Die Matroska Definition selber würde es direkt Unterstützen auch die Liedtexte (und deren Übersetzungen), als Untertiteldateien zeitsynchronisiert mit dem Audiosignal zu hinterlegen.
- Bestehende BWF basierte AIPs von anderen Audioarchiven können durch Umverpackung ohne Änderung der Inhaltsinformation migriert werden [NES13] Seite 88

Der Hauptgrund ist aber das aktuell laufende Projekt der Internet Engineering Task Force IETF mit dem Titel *Codec Encoding for LossLess Archiving and Realtime transmission (cellar):*

> *The preservation of audiovisual materials faces challenges from technological obsolescence, analog media deterioration, and the use of proprietary formats that lack formal open standards. While obsolescence and material degradation are widely addressed, the standardization of open, transparent, self-descriptive, lossless formats remains an important mission to be undertaken by the open source community. [...]*
> *Using existing work done by the development communities of Matroska, FFV1, and FLAC, the Working Group will formalize specifications for these open and lossless formats.*[25]

[25] Internet Engineering Task Force (IETF) cellar Projekt: https://datatracker.ietf.org/wg/cellar/charter/

10.3 Vorschlag für die Systemarchitektur

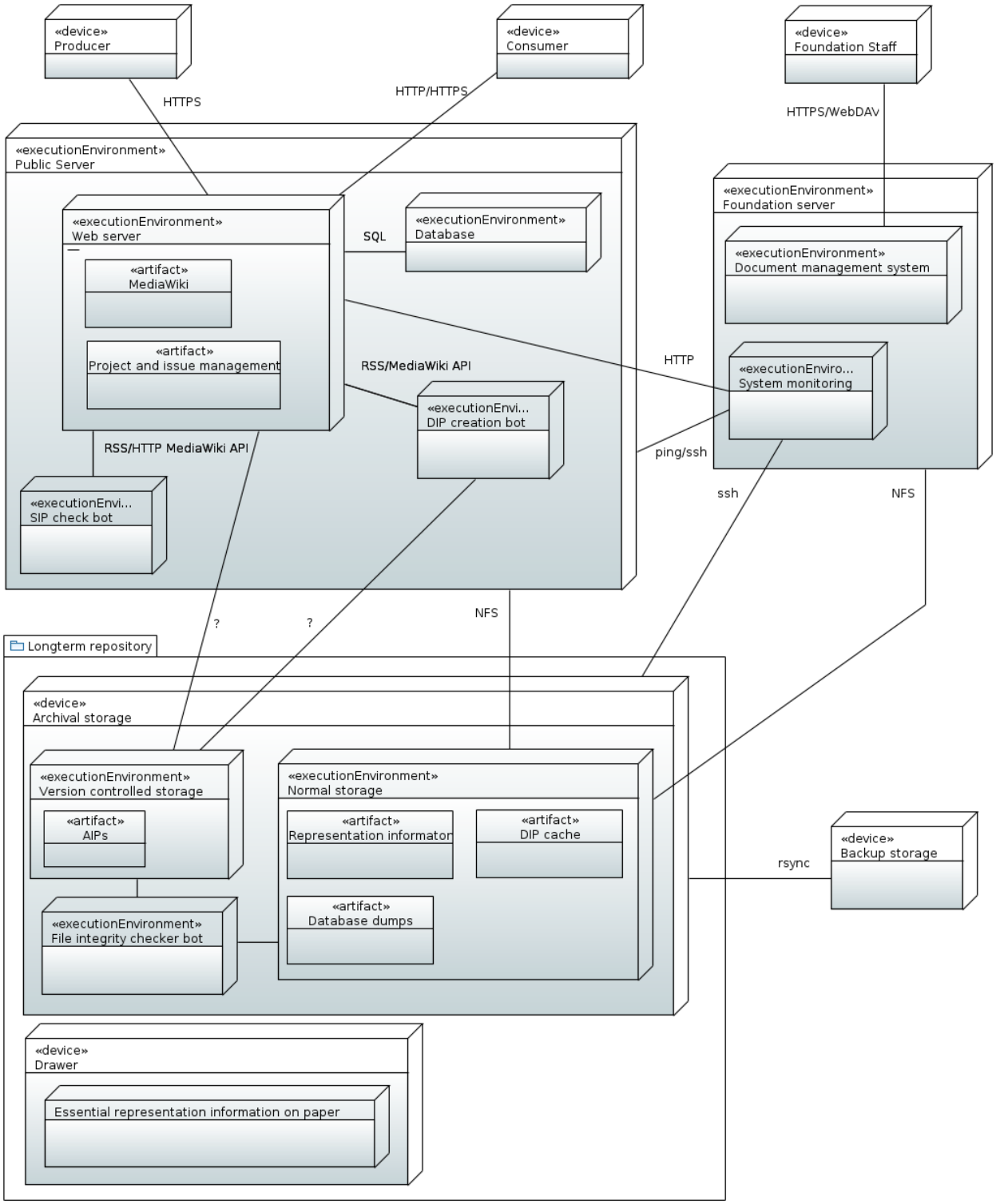

Abbildung 23: Vorgeschlagene Systemarchitektur zur Langzeiterhaltung im Public Domain Projekt

10.3.1 Archivspeicher (Archival storage)

Der Archivspeicher besteht aus zwei Teilen: Einem Gentoo GNU/Linux[26] basierten Server der die magnetischen Festplatten verwaltet, den Speicherplatz im Netzwerk zur Verfügung stellt und die Daten regelmässig auf undokumentierte Veränderungen prüft. Auf dem Server werden die AIPs in einem versionsverwalteten Speichersystem abgelegt. Dieses System stellt sicher, dass alle Änderungen an AIPs aufgezeichnet werden (Wer hat wann was geändert) und speichert von jeder AIP Version die Prüfsumme, um unerwünschte Veränderungen erkennen zu können. Dieses System muss evaluiert werden.

Auf dem Server werden in digitaler Form Repräsentationsinformationen zu den eingesetzten Dateiformaten, Dateisystem, Metadatenstandards und Vokabulare manuell abgelegt. Ergänzt wird die Repräsentationsinformation durch alle Quellcodes der auf dem Server vorhandenen Software. Dies geschieht automatisch da Gentoo Linux jede Softwarekomponente als Quellcode herunterlädt und automatisiert kompiliert und installiert. So ist sichergestellt, dass die Repräsentationsinformation zur Software immer vollständig und aktuell ist. Anstatt Gentoo könnte auch FreeBSD eingesetzt werden, um dies zu erreichen.

Der zweite Teil ist eine Schublade die im Serverraum neben dem Gentoo GNU/Linux Server eingebaut wird. Sie enthält in Papierform die essentiellsten Repräsentationsinformationen um die Daten im Server auslesen und interpretieren zu können. Dazu gehören ASCII Tabelle, Zahlensystembeschreibung, Spezifikation der Festplattenschnittstelle, Dateisystem und Dateiformatspezifikationen, Sprachreferenz für die Programmiersprache C und x86 Befehlssatz.

10.3.2 Geographisch getrennte Sicherheitskopie (Backup storage)

Es soll ein weiteres Speichersystem eingerichtet werden oder eine Partnerschaft eingegangen werden um Sicherheitskopien aller Daten an einem getrennten Ort zu lagern. Die Distanz soll genug gross sein, dass Naturkatastrophen wie Erdbeben, Überschwemmungen etc. nur einen Standort betreffen.

10.3.3 Öffentlicher Server (Public server)

Beim öffentlichen Server handelt es sich um den bestehenden Debian GNU/Linux Server wo derzeit Apache2 mit MediaWiki und MySQL als Datenbank läuft. Dieser Server erfüllt jetzt schon die Anforderungen an eine einfache Softwareschnittstelle (API) per MediaWiki API. Die installierte Erweiterung *Semantic Mediawiki* [27] bietet die Möglichkeit SPARQL anfragen zu verarbeiten, dies kann derzeit aber noch nicht aktiviert werden, da die Metadaten des Public Domain Projekts innerhalb des MediaWiki nicht maschinenlesbar sind.

Der öffentliche Server soll um mehrere Komponenten ergänzt werden. Die im Folgenden näher beschreiben sind.

10.3.3.1 Projekt- und Aufgabenverwaltung (Project and issue management)

Das Audit hat gezeigt, dass in den Bereichen Administration und Erhaltungsplanung viele Aufgaben auf das Public Domain Projekt zukommen, wenn es diese Probleme in Angriff nimmt.

Zu den Aufgaben der Stiftung gehört auch ein Teil des Projektmanagements und der Erhaltungsplanung. Diese Aufgaben werden aber auch von den Freiwilligen/Ehrenamtlichen übernommen, darum wird ein Werkzeug dass diese Arbeiten unterstützt auf dem öffentlichen Server implementiert.

Ein Resultat des Audits ist, dass diverse Bereiche schlecht dokumentiert sind oder Definitionen fehlen. Solche öffentlichen Daten werden und sollen auch weiterhin im MediaWiki ergänzt und ausgebaut,

[26] https://gentoo.org/
[27] https://www.semantic-mediawiki.org

insbesondere Beschreibungen zu Prozessen, Systemarchitektur, Definitionen, Anforderungen (Policies).

Den Anforderungen aus dem Audit ist zu entnehmen, dass viele Aufgaben der Administration und der Erhaltungsplanung regelmässig ausgeführt werden müssen und über alle erledigten Aufgaben Berichte vorhanden sein müssen. Zur Unterstützung wird vorgeschlagen ein Issue Management System zu evaluieren, in dem dann alle anstehenden Aufgaben eingepflegt und nachvollziehbar bearbeitet werden können. Das Issue Management System dient parallel auch zur Erfassung von Problemen, Wünschen und Empfehlungen der Nutzer sowie Rückmeldung über Fehler etc. von den Produzenten. Da alle zu bearbeitenden Aufgaben und Anfragen in einem System gesammelt werden, ist eine hohe Sichtbarkeit gewährleistet. Da das System öffentlich ist, ist es für jeden interessierten Freiwilligen möglich zu sehen, was es im Projekt gerade zu tun gibt und welche der anstehenden Aufgaben übernehmen könnte.

10.3.3.2 Auslieferungsprozess (DIP creation bot)

Ein automatisierter Prozess soll alle AIPs in Flac und Ogg/Vorbis Dateien (DIP) umwandeln und über die Weboberfläche verfügbar machen. Dabei sollten die Metadaten soweit wie vom Dateiformat unterstützt vom AIP übernommen werden. Es soll auch ein Link zum Herunterladen des vollständigen AIP zur Verfügung gestellt werden. Auf der Webseite sollen die Prüfsummen der angebotenen Dateien angezeigt werden, damit überprüft werden kann ob die Datei beim Herunterladen unerwünscht verändert wurde.

10.3.3.3 Eingangsprüfung (SIP check bot)

Um sicherzustellen, dass alle AIPs den Vorgaben für die Erfassung von Erhaltungsmetadaten entsprechen, soll ein automatischer Hintergrundprozess (Bot) den Übernahmeprozess unterstützen. Er soll die SIPs, die in Bearbeitung sind, auf Vollständigkeit und formelle Korrektheit prüfen.

Eine genaue Evaluierung der Anforderungen steht noch aus, da diese auch von der Definition des AIP und den eingesetzten Metadatenstandards abhängig ist. Eine Basis für diesen Bot könnte das von der EU geförderte Open Source Projekt Mediaconch darstellen.

> *MediaConch (CONformance CHecking for audiovisual files) is an extensible, open source software project consisting of an implementation checker, policy checker, reporter and fixer that targets preservation-level audiovisual files*
> *Quelle: http://www.preforma-project.eu/mediaconch.html*

10.3.4 Stiftungsserver (Foundation server)

Die internen Dokumente der Schweizerischen Stiftung Public Domain müssen auch versionsverwaltet werden, derzeit sind sie auf verschiedenen Rechnern lokal abgelegt. Da die Mitglieder des Stiftungsrats in verschiedenen Regionen wohnen, soll ein Dokumentenverwaltungssystem (DMS) evaluiert werden, das einen geschützten Zugang und Synchronisation übers Internet erlaubt.

Die Infrastruktur des Public Domain Projekt muss überwacht werden und Unregelmässigkeiten an die betreffenden Stellen gemeldet werden. Dazu ist ein Monitoringsystem zu evaluieren.

Für diese beiden internen Systeme soll ein neuer Debian GNU/Linux Server eingerichtet werden, um diese Dienste logisch vom öffentlichen Bereich zu trennen.

11 Schlussfolgerungen/Fazit

Erreicht wurden die Übergeordneten Ziele der Arbeit, wie das Erarbeiten der theoretischen und methodischen Grundsätze der Langzeiterhaltung und der Vermittlung dieses Wissens innerhalb des Public Domain Projekts. Das Ziel, den aktuellen Stand zu Analysieren, konnte mit Hilfe des Audits auch erreicht werden. Das Ziel eine gute Metrik für Entscheidungen und deren Konsequenzen zu erarbeiten, konnte mit dem CCSDS 652.0-M-1 Audit und dem Erhaltungsplanungsprozess der KOST-CECO ebenfalls erreicht werden. Es ist daraus klar abzulesen, was die fundamentalen Probleme sind, die aktuell die bestehenden Audiowerke gefährden. Die Anforderungsanalyse konnte so direkt aus den ersten beiden Teilen der Projektarbeit abgeleitet werden.

Das Audit hat wertvolle Resultate geliefert, war aber zeitintensiver als angenommen. Die verwendeten Begriffe im Audit mussten sehr genau untersucht werden, um präzise zu verstehen was gefordert wird. Es kamen ein paar Kriterien vor, für die es keine Definition der Begriffe gab und auch Recherchen keine genaue Definition ergaben.

Die Einarbeitung und die Evaluation von geeigneten Metadatenstandards stellte sich als die grösste Herausforderung dieser Arbeit dar. Die Metadatenstandard-Landschaft ist sehr unübersichtlich und es existieren sehr viele Standards. Die Frage nach der Herkunft und der Zielgruppe eines Metadatenstandards konnte oft erst nach längeren Recherchen klar beantwortet werden. Ein Problem war es auch herausfinden, wie verbreitet ein Standard ist. Dazu konnten während dieser Arbeit keine verlässlichen Daten gefunden werden. Bei Memoriav und befreundeten Experten wurde angefragt, welche Standards üblicherweise verwendet werden. Die Antworten waren nicht sehr hilfreich, da jeweils andere Standards genannt wurden und auch immer welche dabei waren, die in der eigenen Recherche noch nicht angetroffen wurden. Das hohe Ziel der weltweit vernetzten, maschineninterpretierbaren, interoperablen Metadaten scheint durch diese Vielzahl von Standards noch in weiter Ferne zu liegen.

Noch ist nicht sicher, ob die vorgeschlagene Kombination aus DublinCore und PREMIS tatsächlich alle benötigen Erhaltungsmetadaten repräsentieren kann. Wenn während der Implementierungsphase Lücken festgestellt werden, werden weitere Ontologien evaluiert werden müssen um diese zu füllen.

Während dieser Projektarbeit konnte noch nicht am Datenexport gearbeitet werden, weil zum Einen die Zeit dazu nicht mehr vorhanden war und zum Anderen weil die Anforderungen der weiterverarbeitenden Systeme noch nicht bekannt waren. Auch die fehlende Definition der vorgesehenen Zielgruppen war ein Hinderungsgrund.

Die vorgeschlagene Systemarchitektur, wenn sie so umgesetzt wird, wird das Public Domain Projekt ein grosses Stück näher an das Ziel bringen, ein vertrauenswürdiges digitales Langzeitarchiv zu sein.

12 Abbildungsverzeichnis

13 Tabellenverzeichnis

14 Literaturverzeichnis

CAP08: Priscilla Caplan, What Is Digital Preservation?, Library Technology Reports, https://journals.ala.org/ltr/article/view/4224/4809, 2008

CCSDS652: , AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES, CCSDS, CCSDS 652.0-M-1, http://public.ccsds.org/publications/archive/652x0m1.pdf, 2011

KEI13: Christian Keitel, Astrid Schoger, Vertrauenswürdige digitale Langzeitarchivierung nach DIN 31644, Beuth, 978-3-410-23499-9, , 2013

KOS15: kost-ceco, Preservation Process der KOST, 2015, http://kost-ceco.ch/cms/index.php?id=238,422,0,0,1,0

MEM14: Rudolf Müller, Yves Cirio et al, Empfehlungen zur Erhaltung von Tondokumenten, Memoriav, , http://memoriav.ch/wp-content/uploads/2014/06/empfehlun-gen_ton_de.pdf, 2014

NES08: nestor-Arbeitsgruppe „Vertrauenswürdige Archive - Zertifizierung", Kriterienkatalog Vertrauenswürdige digitale Langzeitarchive - Version 2, nestor, urn:nbn:de:0008-2008021802, http://nbn-resolving.de/urn:nbn:de:0008-2008021802, 2008

NES13: nestor-Arbeitsgruppe OAIS-Übersetzung / Terminologie, Referenzmodell für ein Offenes Archiv-Informations-System - Deutsche Übersetzung, Version 2, nestor, urn:nbn:de:0008-2013082706, http://nbn-resolving.de/urn:nbn:de:0008-2013082706, 2013

OADE14: Sabine Schrimpf, Das OAIS-Modell für die Langzeitarchivierung - Anwendung der ISO 14721 in Bibliotheken und Archiven, Beuth, 978-3-410-23954, , 2014

OAIS12: , REFERENCE MODEL FOR AN OPEN ARCHIVALINFORMATION SYSTEM (OAIS), CCSDS, CCSDS 650.0-M-2, http://public.ccsds.org/publications/archive/650x0m2.pdf, 2012

TC03: Dietrich Schüller et al, Die Bewahrung von Schallaufnahmen - Ethische Aspekte, Prinzipien und Strategien, Internationale Vereinigung der Schall- und audiovisuellen Archive, IASA-TC 03, http://www.iasa-web.org/sites/default/files/downloads/publications/TC03_German.pdf, 2005

TRAC07: Robin L. Dale et al, Trustworthy Repositories Audit & Certification: Criteria and Checklist, Center for Research Libraries and Online Computer Library Center, OCLC No. 85786759, http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf, 2007

15 Lizenz

15.1 Lizenz der verwendeten Bilder

Bei Bildern die von anderen Urhebern stammen ist jeweils die Quelle angegeben. Die Lizenzbedingungen können abweichend sein von der Lizenz dieser Arbeit.

Quelle des Titelbildes: Courtesy of the Recorded Sound Section, MBRS Division, Library of Congress

16 Selbständigkeitserklärung

Ich bestätige, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der im Literaturverzeichnis angegebenen Quellen und Hilfsmittel angefertigt habe. Sämtliche Textstellen, die nicht von mir stammen, sind als Zitate gekennzeichnet und mit dem genauen Hinweis auf ihre Herkunft versehen.

Ort, Datum:

Unterschrift:

17 Anhang

- Report des Audits gemäss CCSDS 652.0-M-1
- Liste der Metadatenfelder im Public Domain Projekt
- Liste der Vorbis Tags
- Linkliste mit Webseiten die für diese Arbeit und für die kommende Masterthesis relevant sind
- Aufgabenstellung Projektarbeit 2

PD:Internal CCSDS 652.0-M-1 audit

From PUBLIC DOMAIN PROJECT

Date of audit report: June 2016

Version of audit report: 1.0

This is the first audit of the public domain project. This audit is the starting point for a long term program to develop and install the required organizational and technical methods to fulfill the requirements for a long term digital archive.

This audit was done according to the recommended practice 652.0-M-1 *AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES* from 2011 published by the Consultative Committee for Space Data Systems (CCSDS). The same committee that published the Reference Model for an Open Archival Information System (OAIS) (http://public.ccsds.org/publications/archive/650x0m2.pdf).

It was clear from the beginning, that this first audit will show a lot of weak points, not addressed problems and essential requirements that are not fulfilled.

Therefor the huge amount of requirements marked as not fulfilled (red) should not lead to the verdict that the public domain project is completely not trustworthy. The existence of this audit is more than many other archives provide as publicly available information source to evaluate the trustworthiness.

The result of this first audit is the fundamental work for the development of requirements for future processes, technical methods and investments. It helps also to manage this development projects as it provides a metric to measure the impact of a proposal.

This audit will be replaced by a more recent audit after the implementation of serious improvements. So it is possible to track the efforts the project invests into the longterm preservation and its trustworthiness.

This audit documentation is structured in a similar way as the CCSDS 652.0-M-1 Recommended Practice document:

- introduction
- overview of audit and certification criteria
- conclusion
- catalog of requirements

Contents

repository that affect the preservation of digital content over time.
- 3.3.5 The repository shall define, collect, track, and appropriately provide its information integrity measurements.
- 3.3.6 The repository shall commit to a regular schedule of self-assessment and external certification.

- 3.4 FINANCIAL SUSTAINABILITY
  - 3.4.1 The repository shall have short- and long-term business planning processes in place to sustain the repository over time.
  - 3.4.2 The repository shall have financial practices and procedures which are transparent, compliant with relevant accounting standards and practices, and audited by third parties in accordance with territorial legal requirements.
  - 3.4.3 The repository shall have an ongoing commitment to analyze and report on financial risk, benefit, investment, and expenditure (including assets, licenses, and liabilities).
- 3.5 CONTRACTS, LICENSES, AND LIABILITIES
  - 3.5.1 The repository shall have and maintain appropriate contracts or deposit agreements for digital materials that it manages, preserves, and/or to which it provides access.
    - 3.5.1.1 The repository shall have contracts or deposit agreements which specify and transfer all necessary preservation rights, and those rights transferred shall be documented.
    - 3.5.1.2 The repository shall have specified all appropriate aspects of acquisition, maintenance, access, and withdrawal in written agreements with depositors and other relevant parties.
    - 3.5.1.3 The repository shall have written policies that indicate when it accepts preservation responsibility for contents of each set of submitted data objects.
    - 3.5.1.4 The repository shall have policies in place to address liability and challenges to ownership/rights.
  - 3.5.2 The repository shall track and manage intellectual property rights and restrictions on use of repository content as required by deposit agreement, contract, or license.

- 4 DIGITAL OBJECT MANAGEMENT
  - 4.1 INGEST: ACQUISITION OF CONTENT
    - 4.1.1 The repository shall identify the Content Information and the Information Properties that the repository will preserve.
      - 4.1.1.1 The repository shall have a procedure(s) for identifying those Information Properties that it will preserve.
      - 4.1.1.2 The repository shall have a record of the Content Information and the Information Properties that it will preserve.
    - 4.1.2 The repository shall clearly specify the information that needs to be associated with specific Content Information at the time of its deposit.
    - 4.1.3 The repository shall have adequate specifications enabling recognition and parsing of the SIPs.
    - 4.1.4 The repository shall have mechanisms to appropriately verify the identity of the Producer of all materials.
    - 4.1.5 The repository shall have an ingest process which verifies each SIP for completeness and correctness.
    - 4.1.6 The repository shall obtain sufficient control over the Digital Objects to preserve them.
    - 4.1.7 The repository shall provide the producer/depositor with

appropriate responses at agreed points during the ingest processes.
- 4.1.8 The repository shall have contemporaneous records of actions and administration processes that are relevant to content acquisition.

- 4.2 INGEST: CREATION OF THE AIP
  - 4.2.1 The repository shall have for each AIP or class of AIPs preserved by the repository an associated definition that is adequate for parsing the AIP and fit for long- term preservation needs.
    - 4.2.1.1 The repository shall be able to identify which definition applies to which AIP.
    - 4.2.1.2 The repository shall have a definition of each AIP that is adequate for long- term preservation, enabling the identification and parsing of all the required components within that AIP.
  - 4.2.2 The repository shall have a description of how AIPs are constructed from SIPs.
  - 4.2.3 The repository shall document the final disposition of all SIPs. In particular the following aspect must be checked.
    - 4.2.3.1 The repository shall follow documented procedures if a SIP is not incorporated into an AIP or discarded and shall indicate why the SIP was not incorporated or discarded.
  - 4.2.4 The repository shall have and use a convention that generates persistent, unique identifiers for all AIPs.
    - 4.2.4.1 The repository shall uniquely identify each AIP within the repository.
      - 4.2.4.1.1 The repository shall have unique identifiers.
      - 4.2.4.1.2 The repository shall assign and maintain persistent identifiers of the AIP and its components so as to be unique within the context of the repository.
      - 4.2.4.1.3 Documentation shall describe any processes used for changes to such identifiers.
      - 4.2.4.1.4 The repository shall be able to provide a complete list of all such identifiers and do spot checks for duplications.
      - 4.2.4.1.5 The system of identifiers shall be adequate to fit the repository's current and foreseeable future requirements such as numbers of objects.
    - 4.2.4.2 The repository shall have a system of reliable linking/resolution services in order to find the uniquely identified object, regardless of its physical location.
  - 4.2.5 The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains. In particular the following aspects must be checked.
    - 4.2.5.1 The repository shall have tools or methods to identify the file type of all submitted Data Objects.
    - 4.2.5.2 The repository shall have tools or methods to determine what Representation Information is necessary to make each Data Object understandable to the Designated Community.
    - 4.2.5.3 The repository shall have access to the requisite Representation Information.
    - 4.2.5.4 The repository shall have tools or methods to ensure that the requisite Representation Information is persistently associated with the relevant Data Objects.

- 4.2.6 The repository shall have documented processes for acquiring Preservation Description Information (PDI) for its associated Content Information and acquire PDI in accordance with the documented processes. In particular the following aspects must be checked.
  - 4.2.6.1 The repository shall have documented processes for acquiring PDI.
  - 4.2.6.2 The repository shall execute its documented processes for acquiring PDI.
  - 4.2.6.3 The repository shall ensure that the PDI is persistently associated with the relevant Content Information.
- 4.2.7 The repository shall ensure that the Content Information of the AIPs is understandable for their Designated Community at the time of creation of the AIP. In particular the following aspects must be checked.
  - 4.2.7.1 Repository shall have a documented process for testing understandability for their Designated Communities of the Content Information of the AIPs at their creation.
  - 4.2.7.2 The repository shall execute the testing process for each class of Content Information of the AIPs.
  - 4.2.7.3 The repository shall bring the Content Information of the AIP up to the required level of understandability if it fails the understandability testing.
- 4.2.8 The repository shall verify each AIP for completeness and correctness at the point it is created.
- 4.2.9 The repository shall provide an independent mechanism for verifying the integrity of the repository collection/content.
- 4.2.10 The repository shall have contemporaneous records of actions and administration processes that are relevant to AIP creation.

- 4.3 PRESERVATION PLANNING
  - 4.3.1 The repository shall have documented preservation strategies relevant to its holdings.
  - 4.3.2 The repository shall have mechanisms in place for monitoring its preservation environment.
    - 4.3.2.1 The repository shall have mechanisms in place for monitoring and notification when Representation Information is inadequate for the Designated Community to understand the data holdings.
  - 4.3.3 The repository shall have mechanisms to change its preservation plans as a result of its monitoring activities.
    - 4.3.3.1 The repository shall have mechanisms for creating, identifying or gathering any extra Representation Information required.
  - 4.3.4 The repository shall provide evidence of the effectiveness of its preservation activities.
- 4.4 AIP PRESERVATION
  - 4.4.1 The repository shall have specifications for how the AIPs are stored down to the bit level.
    - 4.4.1.1 The repository shall preserve the Content Information of AIPs.
    - 4.4.1.2 The repository shall actively monitor the integrity of AIPs.
  - 4.4.2 The repository shall have contemporaneous records of actions and administration processes that are relevant to storage and preservation of the AIPs.

- 4.4.2.1 The repository shall have procedures for all actions taken on AIPs.
- 4.4.2.2 The repository shall be able to demonstrate that any actions taken on AIPs were compliant with the specification of those actions.

- 4.5 INFORMATION MANAGEMENT
  - 4.5.1 The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest.
  - 4.5.2 The repository shall capture or create minimum descriptive information and ensure that it is associated with the AIP.
  - 4.5.3 The repository shall maintain bi-directional linkage between each AIP and its descriptive information.
    - 4.5.3.1 The repository shall maintain the associations between its AIPs and their descriptive information over time.
- 4.6 ACCESS MANAGEMENT
  - 4.6.1 The repository shall comply with Access Policies.
    - 4.6.1.1 The repository shall log and review all access management failures and anomalies.
  - 4.6.2 The repository shall follow policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity.
    - 4.6.2.1 The repository shall record and act upon problem reports about errors in data or responses from users.

- 5 INFRASTRUCTURE AND SECURITY RISK MANAGEMENT
  - 5.1 TECHNICAL INFRASTRUCTURE RISK MANAGEMENT
    - 5.1.1 The repository shall identify and manage the risks to its preservation operations and goals associated with system infrastructure.
      - 5.1.1.1 The repository shall employ technology watches or other technology monitoring notification systems.
        - 5.1.1.1.1 The repository shall have hardware technologies appropriate to the services it provides to its designated communities.
        - 5.1.1.1.2 The repository shall have procedures in place to monitor and receive notifications when hardware technology changes are needed.
        - 5.1.1.1.3 The repository shall have procedures in place to evaluate when changes are needed to current hardware.
        - 5.1.1.1.4 The repository shall have procedures, commitment and funding to replace hardware when evaluation indicates the need to do so.
        - 5.1.1.1.5 The repository shall have software technologies appropriate to the services it provides to its designated communities.
        - 5.1.1.1.6 The repository shall have procedures in place to monitor and receive notifications when software changes are needed.
        - 5.1.1.1.7 The repository shall have procedures in place to evaluate when changes are needed to current software.
        - 5.1.1.1.8 The repository shall have procedures, commitment, and funding to replace software when evaluation indicates the need to do so.
      - 5.1.1.2 The repository shall have adequate hardware and

software support for backup functionality sufficient for preserving the repository content and tracking repository functions.
        - 5.1.1.3 The repository shall have effective mechanisms to detect bit corruption or loss.
            - 5.1.1.3.1 The repository shall record and report to its administration all incidents of data corruption or loss, and steps shall be taken to repair/replace corrupt or lost data.
        - 5.1.1.4 The repository shall have a process to record and react to the availability of new security updates based on a risk-benefit assessment.
        - 5.1.1.5 The repository shall have defined processes for storage media and/or hardware change (e.g., refreshing, migration).
        - 5.1.1.6 The repository shall have identified and documented critical processes that affect its ability to comply with its mandatory responsibilities.
            - 5.1.1.6.1 The repository shall have a documented change management process that identifies changes to critical processes that potentially affect the repository's ability to comply with its mandatory responsibilities.
            - 5.1.1.6.2 The repository shall have a process for testing and evaluating the effect of changes to the repository's critical processes.
    - 5.1.2 The repository shall manage the number and location of copies of all digital objects.
        - 5.1.2.1 The repository shall have mechanisms in place to ensure any/multiple copies of digital objects are synchronized.
- 5.2 SECURITY RISK MANAGEMENT
    - 5.2.1 The repository shall maintain a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant.
    - 5.2.2 The repository shall have implemented controls to adequately address each of the defined security risks.
    - 5.2.3 The repository staff shall have delineated roles, responsibilities, and authorizations related to implementing changes within the system.
    - 5.2.4 The repository shall have suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an offsite copy of the recovery plan(s).

1 OVERVIEW OF AUDIT AND CERTIFICATION CRITERIA

1.1 A TRUSTWORTHY DIGITAL REPOSITORY

Definition of a trustworthy digital repository as given in the CCSDS 652.0-M-1 Recommended Practice document:

A trustworthy digital repository will understand threats to and risks within its systems. Constant monitoring, planning, and maintenance, as well as conscious

actions and strategy implementation will be required of repositories to carry out their mission of digital preservation. All of these present an expensive, complex undertaking that depositors, stakeholders, funders, the Designated Community, and other digital repositories will need to rely on in the greater collaborative digital preservation environment that is required to preserve the vast amounts of digital information generated now and into the future.

1.2 DEFINITIONS

Each requirement is marked with a color, to show its status of fulfillment:

- Requirements fulfilled
- Minor requirements are not fulfilled
- Essential requirements not fulfilled

These definitions from the original audit document all apply to this internal audit:

- TERMINOLOGY
- Glossary
- CONVENTIONS

For a better understanding some paragraphs of the CCSDS 652.0-M-1 Recommended Practice are reproduced here.

1.2.1 CONFORMANCE

Original text: *An archive that conforms to this Recommended Practice shall have satisfied the auditor on* each of the requirements.

1.2.2 EVIDENCE

Each metric in the Recommended Practice has associated with it informative text under the heading *Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement* providing examples of the evidence which might be examined to test whether the repository satisfies the metric. These examples are illustrative rather than prescriptive, and the lists of possible evidence are not exhaustive.

1.2.3 NOMENCLATURE

The following conventions apply for the normative specifications in this Recommended Practice:

a) the words ‘shall’ and ‘must’ imply a binding and verifiable specification;
b) the word ‘should’ implies an optional, but desirable, specification;
c) the word ‘may’ implies an optional specification;
d) the words ‘is’, ‘are’, and ‘will’ imply statements of fact.

1.2.4 ACRONYMS AND ABBREVIATIONS

AIP	Archival Information Package (defined in reference [1])
CCSDS	Consultative Committee for Space Data Systems

DEDSL	Data Entity Specification Language
DIP	Dissemination Information Package (defined in reference [1])
FITS	Flexible Image Transport System
GIS	Geographic Information System
ISO	International Organization for Standardization
OAIS	Open Archival Information System (see reference [1])
PDI	Preservation Description Information (defined in reference [1])
SIP	Submission Information Package (defined in reference [1])
TEI	Text Encoding Initiative
UML	Unified Modeling Language
XML	Extensible Markup Language

1.3 REFERENCES

[1] Reference Model for an Open Archival Information System (OAIS) (http://public.ccsds.org/publications/archive/650x0m2.pdf).

For convenience the full text of the recommended practice CCSDS 652.0-M-1 AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES is readable on this wiki page: PD:CCSDS_652.0-M-1. Every requirement is directly linked to the corresponding explanation in the CCSDS 652.0-M-1 Recommended Practice.

The original document is published on the CCSDS Website: CCSDS Recommended Practices (Magenta Books) (http://public.ccsds.org/publications/MagentaBooks.aspx)

2 CONCLUSION AND FIELDS OF NON CONFORMANCE

2.1 OVERVIEW

Of the 108 normative metrics the final status is the following:

Metrics with all requirements fulfilled (green): 16
Metrics where Minor requirements are not fulfilled (orange): 15
Metrics with essential requirements not fulfilled (red): 77

2.2 FIELDS OF NON CONFORMANCE

2.2.1 ESSENTIAL DEFINITIONS

In the project and between the project members there is a mutual understanding of the designated communities but it is not precisely defined and therefor the knowledge base of these communities is not known.

The same is true for the definition of the content information that has to be preserved.

2.2.2 REPRESENTATION INFORMATION

An example of an underdeveloped area is the field of representation information, because the awareness of the underlying problems and the concepts to handle them was missing in the project at the beginning of this audit. The consequent use of open standards and open source software makes it a bit less critical. But because any representation information is missing, it still creates a large long term risk for the repository.

This whole topic has to be addressed in the near future.

2.2.3 MANAGEMENT AND PRESERVATION PLANNING

The area of management tasks, strategic planning, development of policies and tracking is also underdeveloped. Also, there is no risk assessment installed and consequently there are no processes to observe the technical and legal environment of the repository on a regular basis.

Also missing is a system to plan, manage and track work packages, milestones, issues etc. to support the further development of management and preservation planning.

Furthermore a system which enables end users and producers to submit feedback and where submitters can observe the reactions and actions in response to their feedback is missing.

2.2.4 DIGITAL OBJECT MANAGEMENT

It was already known that the handling of the digital objects in the repository has its risks that have not been addressed yet.

The digital objects are at risk because there is no system to prevent unintended deletion of objects, there is no off-site backup and there is no system and associated monitoring to guarantee the bit-level correctness of the digital objects now and in the future.

Also the system to create identifiers for AIPs is not documented and is not ideal for the scalability of the repository.

2.3 CONCLUSION

As it was expected a lot of requirements are not fulfilled. However, the value of this audit is high because it detected very underdeveloped areas inside the project and as such raises the awareness of these problems. Twelve issues them can easily be fixed by documenting what is currently implemented.

It is of high importance to fix the lack of the essential definitions about content information and designated community.

A large field to work on are the regular maintenance and observation tasks of the management and preservation planning. They have to be defined, documented, executed and reviewed.

On the technical side, the completely missing representation information is a substantial shortcoming for a longterm repository. If this information is collected in the near future and maintained thereafter, the real risk of losing understandability is relatively low.

3 ORGANIZATIONAL INFRASTRUCTURE

With this chapter the catalog of requirements starts. Every requirement is explained in the CCSDS 652.0-M-1 document, this explanation can be reached directly by clicking on the heading of the requirement.

3.1 GOVERNANCE AND ORGANIZATIONAL VIABILITY

3.1.1 The repository shall have a mission statement that reflects a commitment to the preservation of, long term retention of, management of, and access to digital information.

Requirements fulfilled

Bylaws §2 of the Swiss Foundation Public Domain

3.1.2 The repository shall have a Preservation Strategic Plan that defines the approach the repository will take in the long-term support of its mission.

Essential requirements not fulfilled

3.1.2.1 The repository shall have an appropriate succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.

Essential requirements not fulfilled

There is no succession plan to address the case the repository ceases to operate or the governing or funding institution substantially changes its scope.

Escrow arrangements are not needed because of the consequent use of free and open source software.

3.1.2.2 The repository shall monitor its organizational environment to determine when to execute its succession plan, contingency plans, and/or escrow arrangements.

Minor requirements are not fulfilled

Financial monitoring is done every year in retrospect to fulfill the accounting requirements for a charity foundation in Switzerland. This includes an external audit by certified layers and is supervised by the Eidgenössische Stiftungsaufsicht (https://www.edi.admin.ch/edi/de/home/fachstellen/eidgenoessische-

stiftungsaufsicht.html).

Monitoring the organizational environment is not in place and fiscal planning is underdeveloped.

3.1.3 The repository shall have a Collection Policy or other document that specifies the type of information it will preserve, retain, manage, and provide access to.

Requirements fulfilled

Bylaws §2 of the Swiss Foundation Public Domain

3.2 ORGANIZATIONAL STRUCTURE AND STAFFING

3.2.1 The repository shall have identified and established the duties that it needs to perform and shall have appointed staff with adequate skills and experience to fulfill these duties.

Essential requirements not fulfilled

3.2.1.1 The repository shall have identified and established the duties that it needs to perform.

Essential requirements not fulfilled

3.2.1.2 The repository shall have the appropriate number of staff to support all functions and services.

Essential requirements not fulfilled

3.2.1.3 The repository shall have in place an active professional development program that provides staff with skills and expertise development opportunities.

Essential requirements not fulfilled

3.3 PROCEDURAL ACCOUNTABILITY AND PRESERVATION POLICY FRAMEWORK

3.3.1 The repository shall have defined its Designated Community and associated knowledge base(s) and shall have these definitions appropriately accessible.

Essential requirements not fulfilled

3.3.2 The repository shall have Preservation Policies in place to ensure its Preservation Strategic Plan will be met.

Essential requirements not fulfilled

3.3.2.1 The repository shall have mechanisms for review, update, and ongoing development of its Preservation Policies as the repository grows and as technology and community practice evolve.

Essential requirements not fulfilled

3.3.3 The repository shall have a documented history of the changes to its operations,

Essential requirements not fulfilled

3.3.4 The repository shall commit to transparency and accountability in all actions supporting the operation and management of the repository that affect the preservation of digital content over time.

Minor requirements are not fulfilled

Reports of financial and technical audits:
Publications of the Foundation do not include yet financial reports because the foundation is quite new and the first report covering 2014 is just finished. But it is planned to publish them.

The first technical audit is the one that is presented in this document. It is published publicly on this page: Internal Audit (CCSDS_652.0-M-1)

Disclosure of governance documents:
As can be seen from other requirements there are no governance documents yet, so there is nothing to be publicly available.

Contracts and agreements with providers of funding and critical services:
They are not publicly available yet.

3.3.5 The repository shall define, collect, track, and appropriately provide its information integrity measurements.

Essential requirements not fulfilled

3.3.6 The repository shall commit to a regular schedule of self-assessment and external certification.

Essential requirements not fulfilled

3.4 FINANCIAL SUSTAINABILITY

3.4.1 The repository shall have short- and long-term business planning processes in place to sustain the repository over time.

Essential requirements not fulfilled

3.4.2 The repository shall have financial practices and procedures which are transparent, compliant with relevant accounting standards and practices,

and audited by third parties in accordance with territorial legal requirements.

Requirements fulfilled

The auditor has witnessed audited annual financial statements for the year 2014 and 2015. As shown above, these statements are not publicly available which should be changed.

3.4.3 The repository shall have an ongoing commitment to analyze and report on financial risk, benefit, investment, and expenditure (including assets, licenses, and liabilities).

Essential requirements not fulfilled

3.5 CONTRACTS, LICENSES, AND LIABILITIES

3.5.1 The repository shall have and maintain appropriate contracts or deposit agreements for digital materials that it manages, preserves, and/or to which it provides access.

Essential requirements not fulfilled

3.5.1.1 The repository shall have contracts or deposit agreements which specify and transfer all necessary preservation rights, and those rights transferred shall be documented.

Essential requirements not fulfilled

3.5.1.2 The repository shall have specified all appropriate aspects of acquisition, maintenance, access, and withdrawal in written agreements with depositors and other relevant parties.

Essential requirements not fulfilled

3.5.1.3 The repository shall have written policies that indicate when it accepts preservation responsibility for contents of each set of submitted data objects.

Essential requirements not fulfilled

3.5.1.4 The repository shall have policies in place to address liability and challenges to ownership/rights.

Requirements fulfilled

3.5.2 The repository shall track and manage intellectual property rights and restrictions on use of repository content as required by deposit agreement, contract, or license.

Requirements fulfilled

The goal of the Public Domain Project is to make accessible digitized audio content.

This requires a thoroughly checking of the intellectual property rights of each work.

The result of this effort can be seen on every detail information page like this example (Gramophone-14678-b45142). Every work includes information about the copyright status in Switzerland, the European Union and the United States including the year when the work enters the public domain. With this information it is possible to track once a year which works entered the public domain (Relevant is only the year, so once a year is enough) and can be made accessible.

4 DIGITAL OBJECT MANAGEMENT

4.1 INGEST: ACQUISITION OF CONTENT

4.1.1 The repository shall identify the Content Information and the Information Properties that the repository will preserve.

Essential requirements not fulfilled

4.1.1.1 The repository shall have a procedure(s) for identifying those Information Properties that it will preserve.

Essential requirements not fulfilled

4.1.1.2 The repository shall have a record of the Content Information and the Information Properties that it will preserve.

Essential requirements not fulfilled

4.1.2 The repository shall clearly specify the information that needs to be associated with specific Content Information at the time of its deposit.

Minor requirements are not fulfilled

The wiki template Audio_fileshows all needed information. But there is limited documentation about it and how to handle it. The section on references is appropriate but the problems start with date fields because there is no date format specified. Also problematic is, that there is no information about the vocabulary to use, should it be a controlled vocabulary (which one) is it free, are there different requirements for each field?

4.1.3 The repository shall have adequate specifications enabling recognition and parsing of the SIPs.

Essential requirements not fulfilled

4.1.4 The repository shall have mechanisms to appropriately verify the identity of the Producer of all materials.

Minor requirements are not fulfilled

To upload AIPs (Flac files) to the archival storage the FTP protocol is used with user authentication. The user name of the person who uploaded a certain file is visible on

the file system of the archival storage as the UNIX owner of the file. This information is not visible and therefor not verifiable by the designated community.

The history of the associated PDI and the identity of its creator is publicly visible and can be verified by everyone. The PDI is stored and edited in wiki pages and MediaWiki requires a password protected user login to edit this information.

4.1.5 The repository shall have an ingest process which verifies each SIP for completeness and correctness.

Essential requirements not fulfilled

4.1.6 The repository shall obtain sufficient control over the Digital Objects to preserve them.

4.1.7 The repository shall provide the producer/depositor with appropriate responses at agreed points during the ingest processes.

Requirements fulfilled

With the depositors there are no agreed points where responses are necessary. But it is possible to get a lot of information about ingested objects by the category listing (each depositor has it's own category to list all his objects), the recent changes page of the wiki and other ways (eg. watchlists). Reports about the ingestion process are done usually annually for the financial supporters.

4.1.8 The repository shall have contemporaneous records of actions and administration processes that are relevant to content acquisition.

Essential requirements not fulfilled

4.2 INGEST: CREATION OF THE AIP

4.2.1 The repository shall have for each AIP or class of AIPs preserved by the repository an associated definition that is adequate for parsing the AIP and fit for long- term preservation needs.

Essential requirements not fulfilled

4.2.1.1 The repository shall be able to identify which definition applies to which AIP.

Essential requirements not fulfilled

4.2.1.2 The repository shall have a definition of each AIP that is adequate for long- term preservation, enabling the identification and parsing of all the required components within that AIP.

Essential requirements not fulfilled

4.2.2 The repository shall have a description of how AIPs are constructed from SIPs.

Essential requirements not fulfilled

The process is not documented but essentially the SIP is the Flac file that is uploaded to the storage server and forms together with the detailed description in the wiki the AIP. So the AIP consists of the wiki page and the linked Flac file.

4.2.3 The repository shall document the final disposition of all SIPs. In particular the following aspect must be checked.

4.2.3.1 The repository shall follow documented procedures if a SIP is not incorporated into an AIP or discarded and shall indicate why the SIP was not incorporated or discarded.

Minor requirements are not fulfilled

4.2.4 The repository shall have and use a convention that generates persistent, unique identifiers for all AIPs.

Essential requirements not fulfilled

Each AIP is identified by a string composed in the following way: <Label>-<Catalog number>-<Order number>

Example:

- Label: Homocord
- Catalog number: B 367
- Order number: M 17234

This results in the URL for the detailed information page: http://pool.publicdomainproject.org/index.php/Homocord-b367-m17234

And the according Flac file name is: homocord-b367-m17234.flac (http://pool.publicdomainproject.org/audio/flac/genres/religious_music/choral/nebe-quartett/homocord-b367-m17234.flac)

4.2.4.1 The repository shall uniquely identify each AIP within the repository.

Requirements fulfilled

4.2.4.1.1 The repository shall have unique identifiers.

Requirements fulfilled

Given that there are no conflicting catalog/order numbers used by a label. This is unlikely but it could happen.

4.2.4.1.2 The repository shall assign and maintain persistent identifiers of the AIP and its components so as to be unique within the context of the repository.

Requirements fulfilled

The described naming scheme is unique in the context of 78rpm records which are the only informations that are currently preserved.

4.2.4.1.3 Documentation shall describe any processes used for changes to such identifiers.

Essential requirements not fulfilled

There is no documentation.

4.2.4.1.4 The repository shall be able to provide a complete list of all such identifiers and do spot checks for duplications.

Minor requirements are not fulfilled

A complete list of all used identifiers is accessible via the category listing Audio file licenses.

There is no automated way to check for duplication. In the wiki it should not be possible to generate duplicates because the page names would create a conflict. There can be only one page with a certain name because no hierarchy is in use. But on the storage server it would be possible to accidentally create duplicates because of the manual upload process and the hierarchical organization (Folder structure by genre/artist).

4.2.4.1.5 The system of identifiers shall be adequate to fit the repository's current and foreseeable future requirements such as numbers of objects.

Essential requirements not fulfilled

It is obvious that this naming scheme is dependent on the naming of the collected items and is tailored to released records. The result are several problems:

- It is unclear how to handle unreleased records (No order/catalog number)
- The archive is open for other recording formats like cylinders, open reel tape and even motion pictures where this naming scheme is not usable
- The naming scheme does not describe how to handle retouched versions (clean master) of the raw digitization (master) where both have to be searchable, distinguishable and accessible

4.2.4.2 The repository shall have a system of reliable linking/resolution services in order to find the uniquely identified object, regardless of its physical location.

Minor requirements are not fulfilled

The naming convention in use is suitable to meet this requirement if only shellac records are archived. Problematic is the missing documentation.

4.2.5 The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains. In particular the following aspects must be checked.

Essential requirements not fulfilled

4.2.5.1 The repository shall have tools or methods to identify the file type of all submitted Data Objects.

Requirements fulfilled

The Unix tool *file* and other more format specific tools are available on the servers.

4.2.5.2 The repository shall have tools or methods to determine what Representation Information is necessary to make each Data Object understandable to the Designated Community.

Essential requirements not fulfilled

No tools or methods in use.

4.2.5.3 The repository shall have access to the requisite Representation Information.

Requirements fulfilled

Due to the fact that the Public Domain Project only uses Free and Open Source Software (FOSS) access to all requisite Representation Information is guarantied.

4.2.5.4 The repository shall have tools or methods to ensure that the requisite Representation Information is persistently associated with the relevant Data Objects.

Essential requirements not fulfilled

This strong requirement is not fulfilled.

4.2.6 The repository shall have documented processes for acquiring Preservation Description Information (PDI) for its associated Content Information and acquire PDI in accordance with the documented processes. In particular the following aspects must be checked.

Essential requirements not fulfilled

4.2.6.1 The repository shall have documented processes for acquiring PDI.

Essential requirements not fulfilled

4.2.6.2 The repository shall execute its documented processes for acquiring PDI.

Essential requirements not fulfilled

4.2.6.3 The repository shall ensure that the PDI is persistently associated with the relevant Content Information.

Requirements fulfilled

At the moment the PDI is stored inside the MediaWiki and is permanently linked to the audio file (Which does not contain PDI). Both the file name of the content information and the wiki page use the same naming scheme so the association is obvious.

4.2.7 The repository shall ensure that the Content Information of the AIPs is understandable for their Designated Community at the time of creation of the AIP. In particular the following aspects must be checked.

Essential requirements not fulfilled

4.2.7.1 Repository shall have a documented process for testing understandability for their Designated Communities of the Content Information of the AIPs at their creation.

Essential requirements not fulfilled

4.2.7.2 The repository shall execute the testing process for each class of Content Information of the AIPs.

Essential requirements not fulfilled

4.2.7.3 The repository shall bring the Content Information of the AIP up to the required level of understandability if it fails the understandability testing.

Essential requirements not fulfilled

4.2.8 The repository shall verify each AIP for completeness and correctness at the point it is created.

Essential requirements not fulfilled

There is no verification process in use. Essentially the SIP is created by the same person that will ingest it and creates the AIP. For example there is no four-eyes principle in use.

4.2.9 The repository shall provide an independent mechanism for verifying the integrity of the repository collection/content.

Essential requirements not fulfilled

4.2.10 The repository shall have contemporaneous records of actions and administration processes that are relevant to AIP creation.

Minor requirements are not fulfilled

For the PDI there the records of actions are automatically captured. Every change on a wiki page is logged, the difference to the previous version can be inspected and the old version can be restored if needed. Here is an example how the version history looks like: Version history of Columbia-a3996-81215 (http://pool.publicdomainproject.org/index.php?title=Columbia-a3996-81215&action=history)

But there is no such thing or other processes for the content information (the Flac files) to capture records of actions.

4.3 PRESERVATION PLANNING

4.3.1 The repository shall have documented preservation strategies relevant to its holdings.

Essential requirements not fulfilled

There are no documented preservation strategies.

4.3.2 The repository shall have mechanisms in place for monitoring its preservation environment.

Essential requirements not fulfilled

There are no formal mechanisms for monitoring the preservation environment. But the active people in the project are in regular contact with groups of the designated communities. This is done by attending conferences, assemblies, regular meetings of user groups. Also the recommendations on formats and media published by the associations of archives or libraries are observed in a informal way.

4.3.2.1 The repository shall have mechanisms in place for monitoring and notification when Representation Information is inadequate for the Designated Community to understand the data holdings.

Essential requirements not fulfilled

4.3.3 The repository shall have mechanisms to change its preservation plans as a result of its monitoring activities.

Essential requirements not fulfilled

This and the next requirement fail because there are no monitoring activities in place on which a reaction could be defined.

4.3.3.1 The repository shall have mechanisms for creating, identifying or gathering any extra Representation Information required.

Essential requirements not fulfilled

4.3.4 The repository shall provide evidence of the effectiveness of its preservation activities.

Essential requirements not fulfilled

4.4 AIP PRESERVATION

4.4.1 The repository shall have specifications for how the AIPs are stored down to the bit level.

Minor requirements are not fulfilled

All file formats used for AIPs or other relevant information are well documented open standards down to the bit level. The representation information is not available locally and it's not linked to the AIPs.

4.4.1.1 The repository shall preserve the Content Information of AIPs.

Essential requirements not fulfilled

No documented work flows.

4.4.1.2 The repository shall actively monitor the integrity of AIPs.

Essential requirements not fulfilled

4.4.2 The repository shall have contemporaneous records of actions and administration processes that are relevant to storage and preservation of the AIPs.

Essential requirements not fulfilled

4.4.2.1 The repository shall have procedures for all actions taken on AIPs.

Essential requirements not fulfilled

4.4.2.2 The repository shall be able to demonstrate that any actions taken on AIPs were compliant with the specification of those actions.

Essential requirements not fulfilled

4.5 INFORMATION MANAGEMENT

4.5.1 The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest.

Requirements fulfilled

All the descriptive information can be searched by the free text search function of the MediaWiki software. For example if someone is interested in instrumental music it can be found with the search term, according to the metadata attribute *Vocal range* with the value *instrumental*.

Search results for *Vocal range instrumental*

Another option to discover material of interest is by using the category system. Every work is added to several categories like genres, country of origin, creation year, recording formats, digitalization devices etc. The example recording above is in several categories, one is the recording label *Decca Records*. This information can be used to show all recordings of *Decca Records* in the public domain archive:

Category:Decca_Records

4.5.2 The repository shall capture or create minimum descriptive information and ensure that it is associated with the AIP.

Minor requirements are not fulfilled

The minimum information requirements are specified by the *Audio file* template in the wiki. This template acts as the input mask when the SIP is built. The template page also includes the documentation about the usage of this template.

Audio file template:
http://pool.publicdomainproject.org/index.php/Template:Audio_file

The template page also contains the available documentation how to use this template. There is no documentation about the vocabulary that should be used to fill the metadata information.

Responsibility for the procurement of metadata lies in the ingestion process where the SIP is assembled.

Capturing provenience and context information with help of this template is done manually and is done until all needed information is found. The minimum level is determined by the requirement that the public domain project is only allowed to publish works that are in the public domain (copyright free). So at least the information to decide on the legal status of the work must be present. This includes title, all authors and there living dates, first release date and publishing label. Technical metadata is also captured with this template like format of the analog recording, devices used for digitalization, catalog and stamper numbers and track length.

A finished SIP could look like this example: Decca-wa782-kwa5215

Additional to the descriptive information captured with the *Audio file* template the SIP gets also context information by categorization. The public domain project uses a polyhierarchical category tree that contains differentiation between genres, country of origin, creation year, recording formats, digitalization devices etc.

Missing is a documentation on the categorization process for an AIP.

As shown in The repository shall have and use a convention that generates persistent, unique identifiers for all AIPs. there is a naming scheme in use that provides persistent, unique identifiers for all AIPs and descriptive information as long as only shellac records are archived.

There is no detailed process work flow documentation.

There is no system and technical architecture documentation.

4.5.3 The repository shall maintain bi-directional linkage between each AIP and its descriptive information.

Minor requirements are not fulfilled

One direction is from the descriptive information to the AIP. This is achieved by a link on the wiki page withe the descriptive information to the Flac file. It is also achieved by the use of a unique, persistent identifier for each work. This identifier is used to name the wiki page containing the descriptive metadata and also the file name of the Flac file.

Example:

- Descriptive Metadata for Hmv-d1388-08247
- The associated AIP with the file name hmv-d1388-08247.flac (http://pool.publicdomainproject.org/audio/flac/genres/classical/chamber_music/budapest_string_quartet/hmv-d1388-08247.flac)

For the other direction the unique, persistent identifier is used to locate the descriptive metadata in the wiki. This can be done by directly entering the URL *http://pool.publicdomainproject.org/index.php/* and the identifier at the end or by using the search function of the wiki.

Beneficial would be a URL (web link) to the descriptive information inside the Flac metadata tags.

For the physical records this bidirectional linking also holds as the context information (category) in the wiki contains the physical location of the record. In the opposite direction the catalog or stamper number can be used to find its descriptive information as well the number of the container to find the context information about the grouped items.

As with other requirements there is a lack of documentation about the process work flow and technical architecture.

4.5.3.1 The repository shall maintain the associations between its AIPs and their descriptive information over time.

Essential requirements not fulfilled

4.6 ACCESS MANAGEMENT

4.6.1 The repository shall comply with Access Policies.

Requirements fulfilled

The public domain project allows free unlimited access to its collection. So there is no need for user accounts or access management to use the available items.

This is documented for the designated communities on the landing page for the media pool and also on the multi language frequently asked questions (FAQ) page:

From Media pool main page:
Creative works of literature, science and art are subject to copyright law. Works in the public domain are those whose intellectual property rights have expired. With the help of volunteers, our team cleans, cataloged and digitized hundreds of gramophone records. After the clearing of copyrights, free works are available inside our media pool and Wikimedia Commons, compressed in Flac without any loss in quality (24-bit/192kHz).

And further down on the same page:
Permission: distributing, reproducing, streaming, sampling, remixing

From the FAQ:
Question: What I'm allowed to do with the music files?

Answer: There is no restriction. You can for ex. redistribute it, copy it, modify it, use it in your own productions, use it as background music and so on.

Unusual for other archives is the fact, that everybody can contribute to the project by supplying additional metadata, context information or by submitting SIPs to the archive. The project is and should be driven by volunteers as it is the base principle of this (and related) archives.

To be able to do so, a user needs to create a user account and a wiki administrator needs to give writing rights to this user. This is more complicated for a wiki than usual but it had to made that strict because of severe spamming problems. Without more wiki administrators the project is not able to maintain easy writing access and keeping the wiki free from spam.

4.6.1.1 The repository shall log and review all access management failures and anomalies.

Minor requirements are not fulfilled

Related to access management failures are two logging systems. The first are the logging features of the MediaWiki software. This logs can be accessed via the *Special pages* link in the wiki: Logs from the data pool wiki (http://pool.publicdomainproject.org/index.php?title=Special%3ALog&type=&user=&page=&year=&month=-1&tagfilter=)

The second logging system are the logs from the web server application (apache2) and a user front end (piwix) to create statistics and analyze this logs:

- Webserver logs for the media data (http://195.176.247.101/stats/index.php?module=CoreHome&action=index&idSite=1&period=month&date=2016-05-18#?module=Actions&action=menuGetDownloads&idSite=1&period=month&date=2016-05-18)
- Webserver logs for the error pages (http://195.176.247.101/stats/index.php?module=CoreHome&action=index&idSite=1&period=month&date=2016-05-05#?module=Actions&action=menuGetPageTitles&idSite=1&period=month&date=2016-05-05). Status code *404* are *Page not found* errors

Not fulfilled is the requirement that written notes should exist of of reviews undertaken or action taken as a result of reviews.

From the discussion in the CCSDS_652.0-M-1 document one important concern is *such as valid users' being denied access*.

Due to the nature of the public domain project this requirement is fulfilled when the requirement *The repository shall maintain the associations between its AIPs and their descriptive information over time.* is met because if it is possible to access the AIP from the descriptive metadata it is possible for the designated communities to get the AIP too. But this requirement is not yet fulfilled.

4.6.2 The repository shall follow policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity.

Requirements fulfilled

From the discussion section of this requirement: *This requirement is concerned only with the relation between DIPs and the AIPs from which they are derived; elsewhere the link between the originals SIPs and the AIPs is considered.*

The public domain project delivers as DIP directly the Flac file from the archival storage without modification. The designated community is able to check the authenticity of this Flac file because there are CRC and hashes included in the Flac file to detect transmission errors.

Additionally it would be helpful for the designated community to include the hash values of each AIP in the metadata details web page.

4.6.2.1 The repository shall record and act upon problem reports about errors in data or responses from users.

Minor requirements are not fulfilled

The repository acts quickly on problem reports from members of the designated community or from internal staff. Most of the time problems are reported by e-mail and then forwarded to the responsible person.

There are no formal processes for problem reports and the reports from the last years are not archived in a central place where they can be reviewed or tracked if they are solved.

5 INFRASTRUCTURE AND SECURITY RISK MANAGEMENT

5.1 TECHNICAL INFRASTRUCTURE RISK MANAGEMENT

Essential requirements not fulfilled

5.1.1 The repository shall identify and manage the risks to its preservation operations and goals associated with system infrastructure.

Essential requirements not fulfilled

5.1.1.1 The repository shall employ technology watches or other technology monitoring notification systems.

Essential requirements not fulfilled

5.1.1.1.1 The repository shall have hardware technologies appropriate to the services it provides to its designated communities.

Minor requirements are not fulfilled

Maintenance of up-to-date Designated Community technology, expectations, and use profiles; provision of bandwidth adequate to support ingest and use demands; systematic elicitation of feedback regarding hardware and service adequacy; maintenance of a current hardware inventory.

The server hardware for hosting the ingest, search and delivery services where upgraded in spring 2015. There performance is very good compared to the current workload. They are ready to handle many more users.

The archival storage system is still fast enough for the current demands and has also still enough storage capacity for the next time. In the archival storage system there are 50% spare slots for additional hard drives.

The Internet connectivity is a symmetrical 1 Gbit/s connection without traffic limitation. For the current user numbers this is enough to achieve short download times.

This requirement is not fulfilled because there exists no current hardware inventory and there is no procedure or system to ask for and receive feedback from the designated communities.

5.1.1.1.2 The repository shall have procedures in place to monitor and receive notifications when hardware technology changes are needed.

Essential requirements not fulfilled

No written procedures but monitoring systems in place to observe server workloads, memory usage, network traffic and free archival storage space.

5.1.1.1.3 The repository shall have procedures in place to evaluate when changes are needed to current hardware.

Essential requirements not fulfilled

5.1.1.1.4 The repository shall have procedures, commitment and funding to replace hardware when evaluation indicates the need to do so.

Essential requirements not fulfilled

5.1.1.1.5 The repository shall have software technologies appropriate to the services it provides to its designated communities.

Minor requirements are not fulfilled

The examples of ways the repository can demonstrate it is meeting this requirement clearly shows that several requirements have to be met:

Maintenance of up-to-date Designated Community technology, expectations, and use profiles
At the time of writing the expectations of the designated community is moving towards mobile use on smart phones and tablets. The MediaWiki software is not very well suited yet for this devices. This is a known problem and the MediaWiki community is working on this. On desktop and laptop computers the project gives access in a usefull way. The finding aids could be improved for the global community.

The on-line radio streams and the page to access this streams seems to fulfill the needs.

Provision of software systems adequate to support ingest and use demands
Software support for ingest is weak.

Systematic elicitation of feedback regarding software and service adequacy
There is no systematic elicitation of feedback about software topics.

Maintenance of a current software inventory
Software inventory is managed via the package manager *apt* used in Debian GNU/Linux. This covers most of the used software like operating system, common services, web server, on-line radio software etc. MediaWiki and the statistics tool piwix are maintained separately. To help the system administrator MediaWiki provides a inventory of installed extensions and version numbers of them together with the version number of the dependencies: Version number of MediaWiki and its extensions

5.1.1.1.6 The repository shall have procedures in place to monitor and receive notifications when software changes are needed.

Essential requirements not fulfilled

5.1.1.1.7 The repository shall have procedures in place to evaluate when changes are needed to current software.

Essential requirements not fulfilled

5.1.1.1.8 The repository shall have procedures, commitment, and funding to replace software when evaluation indicates the need to do so.

Essential requirements not fulfilled

5.1.1.2 The repository shall have adequate hardware and software support for backup functionality sufficient for preserving the repository content and tracking repository functions.

Essential requirements not fulfilled

5.1.1.3 The repository shall have effective mechanisms to detect bit corruption or loss.

Essential requirements not fulfilled

5.1.1.3.1 The repository shall record and report to its administration all incidents of data corruption or loss, and steps shall be taken to repair/replace corrupt or lost data.

Essential requirements not fulfilled

5.1.1.4 The repository shall have a process to record and react to the availability of new security updates based on a risk-benefit assessment.

Requirements fulfilled

To be informed which security updates are available the public domain project is subscribed to this two mailing lists:

- MediaWiki update and security announcements list (https://lists.wikimedia.org/mailman/listinfo/mediawiki-announce)
- Debian Security announcements (https://lists.debian.org/debian-security-announce/)

Log files from the package manager are available on the server to check what software and patches was installed.

The risk-benefit analysis is done by the Debian security team. This volunteers monitor the newly discovered security problems in the Debian stable systems (GNU/Linux operating system and important application software). They prepare, test and provide patches against this problems and try to make sure that the expected behavior does not change.

For the MediaWiki software the risk-benefit analysis is done by the system administrator. But normally MediaWiki security patches are well tested and if there is any side effect it is documented by the developers: Archive of MediaWiki-announce mails (https://lists.wikimedia.org/pipermail/mediawiki-announce/)

5.1.1.5 The repository shall have defined processes for storage media and/or hardware change (e.g., refreshing, migration).

Essential requirements not fulfilled

5.1.1.6 The repository shall have identified and documented critical processes that affect its ability to comply with its mandatory responsibilities.

Essential requirements not fulfilled

5.1.1.6.1 The repository shall have a documented change management process that identifies changes to critical processes that potentially affect the repository's ability to comply with its mandatory responsibilities.

Essential requirements not fulfilled

5.1.1.6.2 The repository shall have a process for testing and evaluating the effect of changes to the repository's critical processes.

Essential requirements not fulfilled

5.1.2 The repository shall manage the number and location of copies of all digital objects.

Essential requirements not fulfilled

5.1.2.1 The repository shall have mechanisms in place to ensure any/multiple copies of digital objects are synchronized.

Essential requirements not fulfilled

5.2 SECURITY RISK MANAGEMENT

5.2.1 The repository shall maintain a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant.

Essential requirements not fulfilled

5.2.2 The repository shall have implemented controls to adequately address each of the defined security risks.

Essential requirements not fulfilled

5.2.3 The repository staff shall have delineated roles, responsibilities, and authorizations related to implementing changes within the system.

Essential requirements not fulfilled

5.2.4 The repository shall have suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an offsite copy of the recovery plan(s).

Essential requirements not fulfilled

Categories: PD:Administration | Archival science

Audio template

Aus PUBLIC DOMAIN PROJECT MEDIAPOOL

These template for new audio file articles can used for the creation of a new article. It includes the data of all required fields. You only have to compare the data. Check it out an existing article for a deeper view.

1 Audio file information

Image(s) see below (if available)

File:StamperID-Audio template.jpg
150px

Label

This file has no **Label information**.
Please add here the **name of the record label** if available. **Example:** Electrola. Usually printed on the record.

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Cat. no.

This file has no **Catalogue number**.
Please add here the **Catalogue number** if available. **Example:** DB 5527. Usually printed on the label of the record.

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Order number

This file has no **Order number**.
Please add here the **Order number** if available. **Example:** 2EA 4608 usally printed below the catalogue number on the left or right side or in the center area.

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Matrix/StamperID

This file has no **Matrix/StamperID**.
Please add here the **Matrix/StamperID** if available. **Example:** 2EA4608^{I} stamped directly on the black surface in the center area of a 78 rpm record. A cylinder record have usally no stamp of the matrix.

Notify the uploader with: {{subst:add-desc-I|1=Audio

template}}

1st release date 1941

1st recording date 22 December 1940[1]

Coupling date unknown

Cutout date unknown

Place of recording London (United Kingdom)

Description HMV C 3214

Author(s)/Composer(s)

This file has no **Author/Composer**.
Please add the **Author** or **Composer** of this work here. **Example:** Johann Sebastian Bach (1685-1750)

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Lyricist(s)

This file has no **Lyricist**.
Please add the **Lyricist** of this work here.
Example: Johann Sebastian Bach (1685-1750)

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Music arranger(s) none

Conductor(s)

This file has no **Conductor**.
Please add the **Conductor** of this work here.
Example: Fritz Stein (1879-1961)

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Performer(s)

This file has no **Performer**.
Please add the **Performer** of this work here.
Example: Berliner Instrumental-Collegium; Max Strub (1900-1966)

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Vocal range

This file has no **Vocal range**.
Please add here the **vocal range (voice) of the performer** if available. **Example:** Tenor. Check online music databases like **CHARM (http://www.charm.rhul.ac.uk/index.html)** or **VICTOR (http://victor.library.ucsb.edu)**. See also the article about Vocal range inside the encyclopedia. For non-vocal music please write "none" or "instrumental".

Notify the uploader with: {{subst:add-desc-I|1=Audio

template}}

Title/Work

This file has no **Title/Work**.
Please add the **Title** of a classical work here.
Example: Konzert in A-moll für Violine und Streichorchester (BWV 1041). Please do not confuse with **Content/Song** (Songs, Arias etc.)

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Content

This file has no **Content/Song**.
Please add the **Content** of a classical work here (**Example:** Nr.1: 1. Satz: Allegro) or add a

Song (**Example:** O sole mio). Please do not confuse with **Title/Work** (musical composition)

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

Genre(s)

This file has no **Genre**.
Please add the **Genre(s)** of this work here.
Example: Concerto

Notify the uploader with: {{subst:add-desc-I|1=Audio template}}

FLAC	FLAC (http://pool.publicdomainproject.org/audio/flac/genres/opera_terminology/catley_gwen/hmv-c3214-2er433.flac), FLAC (Commons)
Ogg (Vorbis/Theora)	none
PD CH	1 January 1997
PD EU	1 January 2012
PD USA	1 January 1997
PD INT	1 January 2012

2 References

1. ↑ CHARM (http://www.charm.rhul.ac.uk): *Composer: BISHOP, Work: Lo, hear the gentle lark, Performer: Orchestra, Date: 1940-12-22; Catalogue: Gray; CatNum: C3214; Date: 1940-12-22; Label: HMV; Performer: Orchestra; Composer: BISHOP; Title: Lo, hear the gentle lark; Num: 2ER 433; Performer: Gwen Catley, soprano; Conductor: Heward, Leslie*, CSV (http://www.charm.rhul.ac.uk/discography/search/2ER_433.csv;jsessionid=0F79E6762AA74C3C2459ECA10228E827.balancer5) of the record

3 Lizenz

Von «http://pool.publicdomainproject.org/index.php?title=Audio_template&oldid=22716»
Kategorien: Pages with broken file links | HMV | 1941 in music | Operas by Frederic Reynolds | Operas | Henry Bishop | Frederic Reynolds | William Shakespeare | Gwen Catley | George Burrows | Leslie Heward | Comedy of Errors (Reynolds) | Lo, here the gentle lark | PD CH 1997 | PD EU 2012 | PD USA 1997 | PD INT 2012 | PD:EURO-SD30 | FLAC sound files | Audio file licenses | Keith Monks 'Archivist Duo Omni' - RCM Mk. IX | Reloop RP-6000 MK5 s | Martin Osterwalder collection

---

- Diese Seite wurde zuletzt am 26. September 2015 um 23:08 Uhr geändert.
- Diese Seite wurde bisher 615 mal abgerufen.
-

Ogg Vorbis I format specification: comment field and header specification

Overview

The Vorbis text comment header is the second (of three) header packets that begin a Vorbis bitstream. It is meant for short, text comments, not arbitrary metadata; arbitrary metadata belongs in a separate logical bitstream (usually an XML stream type) that provides greater structure and machine parseability.

The comment field is meant to be used much like someone jotting a quick note on the bottom of a CDR. It should be a little information to remember the disc by and explain it to others; a short, to-the-point text note that need not only be a couple words, but isn't going to be more than a short paragraph. The essentials, in other words, whatever they turn out to be, eg:

> "Honest Bob and the Factory-to-Dealer-Incentives, _I'm Still Around_, opening for Moxy Früvous, 1997"

Comment encoding

Structure

The comment header logically is a list of eight-bit-clean vectors; the number of vectors is bounded to 2^32-1 and the length of each vector is limited to 2^32-1 bytes. The vector length is encoded; the vector contents themselves are not null terminated. In addition to the vector list, there is a single vector for vendor name (also 8 bit clean, length encoded in 32 bits). For example, the 1.0 release of libvorbis set the vendor string to "Xiph.Org libVorbis I 20020717".

The comment header is decoded as follows:

```
1) [vendor_length] = read an unsigned integer of 32 bits
2) [vendor_string] = read a UTF-8 vector as [vendor_length] octets
3) [user_comment_list_length] = read an unsigned integer of 32 bits
4) iterate [user_comment_list_length] times {

     5) [length] = read an unsigned integer of 32 bits
     6) this iteration's user comment = read a UTF-8 vector as [length] octets

   }

7) [framing_bit] = read a single bit as boolean
8) if ( [framing_bit] unset or end of packet ) then ERROR
9) done.
```

Content vector format

The comment vectors are structured similarly to a UNIX environment variable. That is, comment fields consist of a field name and a corresponding value and look like:

```
comment[0]="ARTIST=me";
comment[1]="TITLE=the sound of Vorbis";
```

- A case-insensitive field name that may consist of ASCII 0x20 through 0x7D, 0x3D ('=') excluded. ASCII 0x41 through 0x5A inclusive (A-Z) is to be considered equivalent to ASCII 0x61 through 0x7A inclusive (a-z).
- The field name is immediately followed by ASCII 0x3D ('='); this equals sign is used to terminate the field name.
- 0x3D is followed by the 8 bit clean UTF-8 encoded value of the field contents to the end of the field.

Field names

Below is a proposed, minimal list of standard field names with a description of intended use. No single or group of field names is mandatory; a comment header may contain one, all or none of the names in this list.

TITLE
: Track/Work name

VERSION
: The version field may be used to differentiate multiple versions of the same track title in a single collection. (e.g. remix info)

ALBUM
: The collection name to which this track belongs

TRACKNUMBER
: The track number of this piece if part of a specific larger collection or album

ARTIST
: The artist generally considered responsible for the work. In popular music this is usually the performing band or singer. For classical music it would be the composer. For an audio book it would be the author of the original text.

PERFORMER
: The artist(s) who performed the work. In classical music this would be the conductor, orchestra, soloists. In an audio book it would be the actor who did the reading. In popular music this is typically the same as the ARTIST and is omitted.

COPYRIGHT
: Copyright attribution, e.g., '2001 Nobody's Band' or '1999 Jack Moffitt'

LICENSE
: License information, eg, 'All Rights Reserved', 'Any Use Permitted', a URL to a license such as a Creative Commons license ("www.creativecommons.org/blahblah/license.html") or the EFF Open Audio License ('distributed under the terms of the Open Audio License. see http://www.eff.org/IP/Open_licenses /eff_oal.html for details'), etc.

ORGANIZATION
: Name of the organization producing the track (i.e. the 'record label')

DESCRIPTION
: A short text description of the contents

GENRE
: A short text indication of music genre

DATE
: Date the track was recorded

LOCATION
: Location where track was recorded

CONTACT
: Contact information for the creators or distributors of the track. This could be a URL, an email address, the physical address of the producing label.

ISRC
: ISRC number for the track; see the ISRC intro page for more information on ISRC numbers.

Implications

- Field names should not be 'internationalized'; this is a concession to simplicity not an attempt to exclude the majority of the world that doesn't speak English. Field contents, however, use the UTF-8 character encoding to allow easy representation of any language.
- We have the length of the entirety of the field and restrictions on the field name so that the field name is bounded in a known way. Thus we also have the length of the field contents.
- Individual 'vendors' may use non-standard field names within reason. The proper use of comment fields should be clear through context at this point. Abuse will be discouraged.
- There is no vendor-specific prefix to 'nonstandard' field names. Vendors should make some effort to avoid arbitrarily polluting the common namespace. We will generally collect the more useful tags here to help with standardization.
- Field names are not required to be unique (occur once) within a comment header. As an example, assume a track was recorded by three well know artists; the following is permissible, and encouraged:

```
ARTIST=Dizzy Gillespie
ARTIST=Sonny Rollins
ARTIST=Sonny Stitt
```

Encoding

The comment header comprises the entirety of the second bitstream header packet. Unlike the first bitstream header packet, it is not generally the only packet on the second page and may not be restricted to within the second bitstream page. The length of the comment header packet is (practically) unbounded. The comment header packet is not optional; it must be present in the bitstream even if it is effectively empty.

The comment header is encoded as follows (as per Ogg's standard bitstream mapping which renders least-

significant-bit of the word to be coded into the least significant available bit of the current bitstream octet first):

1. Vendor string length (32 bit unsigned quantity specifying number of octets)
2. Vendor string ([vendor string length] octets coded from beginning of string to end of string, not null terminated)
3. Number of comment fields (32 bit unsigned quantity specifying number of fields)
4. Comment field 0 length (if [Number of comment fields]>0; 32 bit unsigned quantity specifying number of octets)
5. Comment field 0 ([Comment field 0 length] octets coded from beginning of string to end of string, not null terminated)
6. Comment field 1 length (if [Number of comment fields]>1...)...

This is actually somewhat easier to describe in code; implementation of the above can be found in vorbis/lib /info.c:_vorbis_pack_comment(),_vorbis_unpack_comment()

CCSDS 650.0-M-2
Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2. June 2012.
This Recommended Practice defines the Reference Model for an Open Archival Information System (OAIS). The current issue includes clarifications to many concepts, in particular, Authenticity with the concept of Transformational Information Property introduced; corrections and improvements in diagrams; addition of Access Rights Information to PDI.
http://public.ccsds.org/publications/archive/650x0m2.pdf

CCSDS 652.0-M-1
Audit and Certification of Trustworthy Digital Repositories. Magenta Book. Issue 1. September 2011.
This Recommended Practice defines an audit and certification process for assessing the trustworthiness of digital repositories.
ISO Equivalent : 16363
http://public.ccsds.org/publications/archive/652x0m1.pdf

CCSDS 652.1-M-2
Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories. Magenta Book. Issue 2. March 2014.
The Recommended Practice for Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories contains binding and verifiable specifications required to be fulfilled by organizations that perform ISO audits for assessing the trustworthiness of digital repositories using CCSDS 652.0-M-1, Audit and Certification of Trustworthy Digital Repositories, and provides the appropriate certification. The current issue has been updated to reflect input from ISO experts.
http://public.ccsds.org/publications/archive/652x1m2.pdf

Willkommen bei nestor, dem deutschen Kompetenznetzwerk zur digitalen Langzeitarchivierung. In nestor arbeiten Bibliotheken, Archive, Museen sowie führende Experten gemeinsam zum Thema Langzeitarchivierung und Langzeitverfügbarkeit digitaler Quellen.
http://www.langzeitarchivierung.de/Subsites/nestor/DE/Home/home_node.html

Dieses Wiki ist die Arbeitsumgebung von nestor, dem deutschen Kompetenznetzwerk zur digitalen Langzeitarchivierung.
Das Wiki ergänzt damit die nestor-Webseite, auf der sich u.a. Veranstaltungshinweise und die nestor-Publikationen finden.
https://wiki.dnb.de/display/NESTOR/Startseite

nestor-Handbuch:
Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Download (8,5 MB)
Dieses Werk ist unter einer Creative Commons-Lizenz lizenziert.
(Namensnennung-Keine kommerzielle Nutzung-Weitergabe unter gleichen Bedingungen 3.0 Unported)
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2010071949

Digital Preservation Defined
“This issue of Library Technology Reports,” Priscilla Caplan notes, “is intended to provide a relatively brief, relatively comprehensive introduction to digital preservation.”
In the February/March 2008 issue of LTR, chapter 1 (“What Is Digital Preservation?”) describes digital preservation in terms of what it is (definitions) and what it does (goals and strategies)
https://journals.ala.org/ltr/issue/view/119

Memoriav Verein zur Erhaltung des audiovisuellen Kulturgutes der Schweiz
Memoriav setzt sich aktiv und nachhaltig für die Erhaltung, die Valorisierung
und die breite Nutzung des audiovisuellen Kulturgutes der Schweiz ein.
Memoriav organisiert ein Netzwerk aller an dieser Aufgabe beteiligten,
verantwortlichen und interessierten Institutionen und Personen.
http://memoriav.ch/

Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (Schweiz):
http://kost-ceco.ch

We often get asked where to start with Semantic Web and RDF. There are probably many ways to do this and on this page you will find the one I found useful for myself. There is a lot of stuff to read so let us start with some basics first:
http://strangelove.netlabs.org/semantic-web-basics/

Resource Description Framework (RDF)
RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.
This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations.
https://www.w3.org/2001/sw/wiki/RDF

RDF 1.1 Primer
This primer is designed to provide the reader with the basic knowledge required to effectively use RDF. It introduces the basic concepts of RDF and shows concrete examples of the use of RDF. Secs. 3-5 can be used as a minimalist introduction into the key elements of RDF.
https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

Linked Data: Evolving the Web into a Global Data Space

This book gives an overview of the principles of Linked Data as well as the Web of Data that has emerged through the application of these principles. The book discusses patterns for publishing Linked Data, describes deployed Linked Data applications and examines their architecture:
http://linkeddatabook.com/editions/1.0/

Early Dublin Core workshops popularized the idea of "core metadata" for simple and generic resource descriptions. The fifteen-element "Dublin Core" achieved wide dissemination as part of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and has been ratified as IETF RFC 5013, ANSI/NISO Standard Z39.85-2007, and ISO Standard 15836:2009.
Starting in 2000, the Dublin Core community focused on "application profiles" -- the idea that metadata records would use Dublin Core together with other specialized vocabularies to meet particular implementation requirements. During that time, the World Wide Web Consortium's work on a generic data model for metadata, the Resource Description Framework (RDF), was maturing. As part of an extended set of DCMI Metadata Terms, Dublin Core became one of most popular vocabularies for use with RDF, more recently in the context of the Linked Data movement.
http://dublincore.org/

MusicBrainz is an open music encyclopedia that collects music metadata and makes it available to the public.
MusicBrainz aims to be:
-The ultimate source of music information by allowing anyone to contribute and releasing the data under open licenses.
-The universal lingua franca for music by providing a reliable and unambiguous form of music identification, enabling both people and machines to have meaningful conversations about music.
Like Wikipedia, MusicBrainz is maintained by a global community of users and we want everyone – including you – to participate and contribute.
https://musicbrainz.org/

Welcome to MusicBrainz! This beginners' guide should get you started on both correcting tags in your digital music and contributing data back to MusicBrainz. If this is your first visit to this page, it might be good to read it all before diving into more advanced topics.
https://musicbrainz.org/doc/Beginners_Guide

The AcousticBrainz project aims to crowd source acoustic information for all music in the world and to make it available to the public. This acoustic information describes the acoustic characteristics of music and includes low-level spectral information and information for genres, moods, keys, scales and much more. The goal of AcousticBrainz is to provide music technology researchers and open source hackers with a massive database of information about music.
https://acousticbrainz.org/

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others. Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.
https://www.wikidata.org/wiki/Wikidata:Main_Page

Photo Metadata
IPTC Photo Metadata sets the industry standard for administrative, descriptive, and copyright information about images.
https://iptc.org/standards/photo-metadata/

The Extensible Metadata Platform (XMP) is an ISO standard, originally created by Adobe Systems Inc., for the creation, processing and interchange of standardized and custom metadata for digital documents and data sets. XMP standardizes a data model, a serialization format and core properties for the definition and processing of extensible metadata. Although metadata can alternatively be stored in a sidecar file, embedding metadata avoids problems that occur when metadata are stored separately. Verwandt mit RDF.
https://en.wikipedia.org/wiki/Extensible_Metadata_Platform

The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability.
http://www.loc.gov/standards/premis/

Digital Preservation Metadata and Improvements to PREMIS in Version 3.0
A DCMI/ASIST Joint Webinar presented by Angela Dappert (Wednesday, May 27, 2015)
http://www.loc.gov/standards/premis/v3/tutorial.html

PRONOM is a resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.
https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

https://github.com/timrdf/prizms/wiki/PRONOM

Tech 3293 (EBUCore) is the flagship of our metadata specifications. Combined with the EBU Class Conceptual Model (CCDM) of simple business objects, it provides the framework for descriptive and technical metadata for use in service orientated architectures and audiovisual ontologies for semantic web and linked data developments. EBUCore is adopted by several broadcasters and is referenced by the UK Digital Production

Tech 3293 (EBUCore) is the flagship of our metadata specifications. Combined with the EBU Class Conceptual Model (CCDM) of simple business objects, it provides the framework for descriptive and technical metadata for use in service orientated architectures and audiovisual ontologies for semantic web and linked data developments. EBUCore is adopted by several broadcasters and is referenced by the UK Digital Production Partnership. It was selected as the best 'core' specifications and integrated in MediaCorp's metadata framework. EBUCore is the foundation for technical and descriptive metadata in FIMS and it is the metadata scheme of reference in the EUScreen project (a European portal on audiovisual public archives counting 12 EBU Members and national archives) which delivers linked data to Europeana. It has been published as AES60 by the Audio Engineering Society. MediaInfo is regularly updated to extract and map metadata to EBUCore from a variety of file formats (currently under the sponsorship of the library of Wales). The Nordiff Group is also using EBUCore for metadata exchange between Nordic countries and the Swiss archive project "Memoriav" uses it for its Memobase.
Audio and metadata experts have defined an extended audio model, which schema has been published in EBUCore. The model is being discussed in ITU, SMPTE and AES.
https://tech.ebu.ch/MetadataEbuCore
https://www.ebu.ch/metadata/ontologies/ebucore/

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.
Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.
https://schema.org/

LMER und UOF
Für eine funktionstüchtige Strategie zur Langzeitarchivierung elektronischer Dokumente ist die Erfassung von geeigneten technischen Metadaten unerlässlich. Die Deutsche Nationalbibliothek hat mit Langzeitarchivierungsmetadaten für elektronische Ressourcen (LMER) dafür ein Schema vorgestellt, welches auf einem Modell der Nationalbibliothek von Neuseeland beruht. Im Rahmen des Projekts kopal wurde auf Basis von LMER und dem Standard METS das Paketformat Universelles Objektformat (UOF) entwickelt für Langzeitarchive gemäß dem OAIS-Referenzmodell.
http://www.dnb.de/DE/Standardisierung/LMER/lmer_node.html

Codec Encoding for LossLess Archiving and Realtime transmission (cellar)
IETF standartiesierung von Flac, FFV1 und Matroska.
https://datatracker.ietf.org/wg/cellar/charter/

FLAC stands out as the fastest and most widely supported lossless audio codec, and the only one that at once is non-proprietary, is unencumbered by patents, has an open-source reference implementation, has a well documented format and API, and has several other independent implementations.
https://xiph.org/flac/index.html

https://xiph.org/flac/id.html

Welcome to the Home of Matroska the extensible, open source, open standard Multimedia container
https://matroska.org/

The FLOSS inventory lists all the Free Libre Open Source Software relevant to Europeana and the cultural heritage world.
http://bgweb.nl/floss/

Sourceforge Categorie: Home & Education / Library Software
https://sourceforge.net/directory/home-education/library/os:linux/

The AcousticBrainz project aims to crowd source acoustic information for all music in the world and to make it available to the public.
We have two clients (graphical and command-line) available for people to use to submit audio features to the AcousticBrainz database. These clients calculate a JSON file containing data about the audio and upload it to AcousticBrainz. We do not upload your audio.
https://acousticbrainz.org/download

MediaConch (CONformance CHecking for audiovisual files) is an extensible, open source software project consisting of an implementation checker, policy checker, reporter and fixer that targets preservation-level audiovisual files
http://www.preforma-project.eu/mediaconch.html

SoX is a cross-platform (Windows, Linux, MacOS X, etc.) command line utility that can convert various formats of computer audio files in to other formats. It can also apply various effects to these sound files, and, as an added bonus, SoX can play and record audio files on most platforms.
http://sox.sourceforge.net/

MediaInfo is a convenient unified display of the most relevant technical and tag data for video and audio files.
https://mediaarea.net/en/MediaInfo

Tools for preservation metadata implementation. This document contains
information about tools (e.g. software, scripts, stylesheets) which
support the implementation of preservation metadata, particularly as
defined in the PREMIS data dictionary.

http://www.loc.gov/standards/premis/tools.html

DAITSS is a digital preservation software application. DAITSS is used
by the Florida Digital Archive (FDA), a long-term preservation
repository service provided by the Florida Virtual Campus.

DAITSS provides automated support for the functions of Submission,
Ingest, Archival Storage, Access, Withdrawal, and Repository
Management. It is architected as a set of RESTful Web Services and
micro-services but enforces strict controls to ensure the integrity and
authenticity of archived content. It implements active preservation
strategies based on format-specific processing including, where
necessary, normalization and forward migration. It is particularly well
suited for materials in text, document, image, audio and video formats.

DAITSS was written for a multi-user environment and supports consortial
as well as institutional preservation repositories.

DAITSS is available for use through a GPLv3 license.
http://daitss.fcla.edu/content/welcome-daitss-website-0
https://github.com/daitss

PREMIS metadata in Archivematica. Archivematica supports the entry of PREMIS rights metadata during transfer or ingest. Rights entered via the GUI interface apply to the entirety of the SIP or transfer.
https://www.archivematica.org/en/docs/archivematica-1.5/user-manual/metadata/premis/#premis-template

Henry - a DSP-driving SPARQL end-point. This server hosts a SPARQL end-point able to perform audio processing tasks to answer a particular query. It builds on top of N3 and Transaction Logic.
http://dbtune.org/henry/

MASTER OF SCIENCE IN ENGINEERING

Bern University of Applied Sciences, Biel
MRU – HuCE
Institute for Human Centered Engineering

Human Interface Technology
Project 2

Student: Christoph Zimmermann
Advisor: Daniel Debrunner
Semester: part time, 3rd semester

Subject Title: **Long time archive for audio works**

Abstract:

The Swiss Foundation Public Domain is responsible for the long time data archive of the volunteer driven Public Domain Project. The volunteers are collecting, digitizing completing metadata and investigating the copyright status of old audio records, mainly 78 rpms (Shellac records).
The goals of this Project 2 are: First to analyze the current state of the project, its work flow and digital data conservation strategy according to established standards in professional archives. Second to develop a strategy to bring the project to a professional level and write the requirements specification for the software tools that are needed for this. Third to evaluate suitable metadata standards and open source tools to fulfill this requirements. And fourth to address the already known most important problems which are:

- The Metadata of the audio files on the website are only human readable and therefor are not accessible for other software/platforms and it is not possible to export them properly, which is essential to migrate the data to other tools.
- The metadata on the website and the metadata embedded in the audio files are not in sync.

So the tools in use need to be adopted/extended and an modified workflow must be implemented to fix this problems. This will enable the full implementation of the developed strategy in the following master thesis.

Literature:

- [Schrim] Das OAIS-Modell für die Langzeitarchivierung, 1st edition 2014, DIN, *Sabine Schrimpf.*
- [Keitel] Vertrauenswürdige digitale Langzeitarchivierung nach DIN 31644, 1st edition 2013, DIN, Christian Keitel, Astrid Schoger.
- [Keyser] Indexing - From thesauri to the Semantic Web, 1st edition 2012, Chandos Publishing, Pierre de Keyser

Project start:	Mo. 22.02.2016
Report delivery:	Fr. 05.08.2016
Presentation date:	Fr. 12.08.2016

Assessment:	15 min presentation plus questions/answers Project report

Date: 24th February 2016