Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

- Institute for Human Centered Engineering HuCE

Ich darf sie begrüssen zu meiner Projektarbeit 2 Präsentation.

Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Long time archive for audio works

Project 2

Preservation of audio works like music, speeches etc. in the digital domain for future generations

Author: Christoph Zimmermann
Adviser: Daniel Debrunner
Cooperation: Schweizerische Stiftung Public Domain

- Institute for Human Centered Engineering HuCE

Das Thema dieser Projektarbeit 2 under der folgenden Master Thesis ist ein Langzeitarchiv für digitale Audiowerke wie Musik, Reden, Hörspiele etc.

Introduction

Structure of this presentation

- Background and motivation
- Basics of digital long time preservation
- Audit process, requirements engineering
- Proposed new system architecture
- Conclusion

Zuerst kurz ein Überblick auf die Struktur dieser Präsentation.

Background and motivation

The Public Domain Project

This project was done in cooperation with the Swiss Foundation Public Domain.

The foundation is operating the volunteer based Public Domain Project. A digital repository for audiovisual cultural heritage to preserve it for future generations.

www.publicdomainproject.org

Dieses Projekt entstand in Zusammenarbeit mit der Schweizerischen Stiftung Public Domain. Diese Betreibt das Public Domain Projekt, in dem ehrenamtliche Helfer analoge Tonträger sammeln, digitalisieren und verfügbar machen. Vergleichbar mit dem Gutenberg Projekt für Bücher.

Background and motivation

Current challenges

- There is the awareness that the project is not meeting the requirements of the field of digital long time preservation
  - The processes in the project have grown into there current form
  - Digital data handling an under developed field
- Wish for a structured approach to plan the next development stages

Den Projektmitgliedern ist bewusst, dass die gewachsenen Prozesse nicht genügen um die Anforderungen an eine seriöse Langzeiterhaltung zu erfüllen.

Darum entstand der Wunsch mit einem strukturierten Vorgehen den aktuellen Stand zu erfassen und daraus die weitere Entwicklung zu planen.

In einem ersten Schritt habe ich mich in das Feld der digitalen Langzeiterhaltung eingearbeitet, dessen Grundlagen ich hier kurz vermittle.

Paradigm shift in preservation

Preservation by analog media conservation

- Every copy has a loss of information
- Focus on preserving the only original

Berner Fachhochschule | Haute école spécialisée bernoise | Bern University of Applied Scienc

Die Erhaltung von digitalen Objekten musste sich zuerst von der Denkweise der analogen Erhaltung lösen:

Jede analoge Kopie weist einen Verlust gegenüber dem Original auf. Darum überhaupt gibt es diese Unterscheideung Original und Kopie.

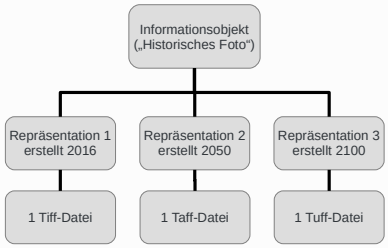
Entsprechend liegt der Fokus auf die aufwändige Erhaltung des **einzigen** Originals.

Source: Museum für Hamburgische Geschichte https://commons.wikimedia.org/wiki/File:MHG_Arbeits , CC BY SA

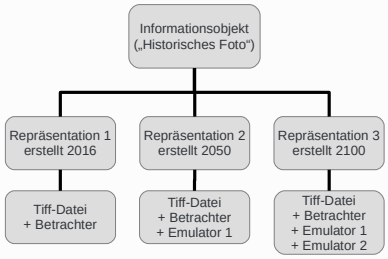
Paradigm shift in preservation

Preservation by digital migration

- Digital copies are equal
- Separation of information and carrier medium
- Focus on reacting to changing environment



Transformation strategy



Emulation strategy

In der digitalen Domäne hingegen können verlustfreie Kopien erstellt werden.

Somit kann die Inhaltsinformation unabhängig vom Speichermedium erhalten bleiben.

Die Inhaltsinformation wird im Laufe der Zeit Migriert. Dazu stehen zwei Strategien, die Transformation und Emulation zur Verfügung.

Simple example: *.txt file

What do we have to preserve to be able to display and understand an ordinary text file (*.txt)?

Aber was muss eigentlich Erhalten werden?

Dazu ein kurzes Beispiel mit einer gewöhnlichen Textdatei.

Simple example: *.txt file

- An ASCII Table (Nowadays a Unicode table)
- Is that enough?

Um eine Textdatei anzeigen zu können
brauchen wir eine ASCII Tabelle.
Reicht diese um die Textdatei zu erhalten?

Simple example: *.txt file

No!

Nein!

Simple example: *.txt file

- Representation information:
  - An ASCII and Unicode table
  - Technical specifications about the file system, storage media, hardware and software interfaces etc. down to the detail level of every bit.
  - Dictionary for the used (human) language
- Metadata about (Preservation information):
  - What, who, why, when, where and the history since it is in the repository

Es sind wesentlich mehr Informationen nötig zu technischen Details bis zum letzten Bit auf dem Datenträger aber auch zur verwendeten Sprache.

Oder können sie als Akademiker heute noch Latein?

Zusätzlich wird ein Satz an Begleitinformationen über Herkunft und Werdegang der Information benötigt.

Zusammengefasst müssen folgende Eigenschaften eines digitalen Objekts erhalten bleiben:

Verfügbarkeit, Identität, Verstehbarkeit, Korrektheit, Lesbarkeit, Darstellbarkeit. Diese Eigenschaften sind auch nötig um die Authentizität sicherzustellen.

Source: Priscilla Caplan, What Is Digital Preservation?, Library Technology Reports, https://journals.a-la.org/ltr/article/view/4224/4809 , 2008

OAIS reference model

Die fachliche und akademische Grundlage der digitalen Langzeiterhaltung ist das OAIS Referenzmodell.

Reference Model For An **O**pen **A**rchival **I**nformation **S**ystem (OAIS)

- THE standard in the field
  - Released as open standard by CCSDS and also as ISO 14721:2012
- Definition of a Long Term Archive:
  [...] 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.
- Definition of Long Term Preservation:
  The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.

Es ist als offener Standard und als ISO Norm publiziert.

Es enthält auch die grundlegenden Begriffsdefinitionen in diesem Bereich.

What means long term?

A period of time long enough for:

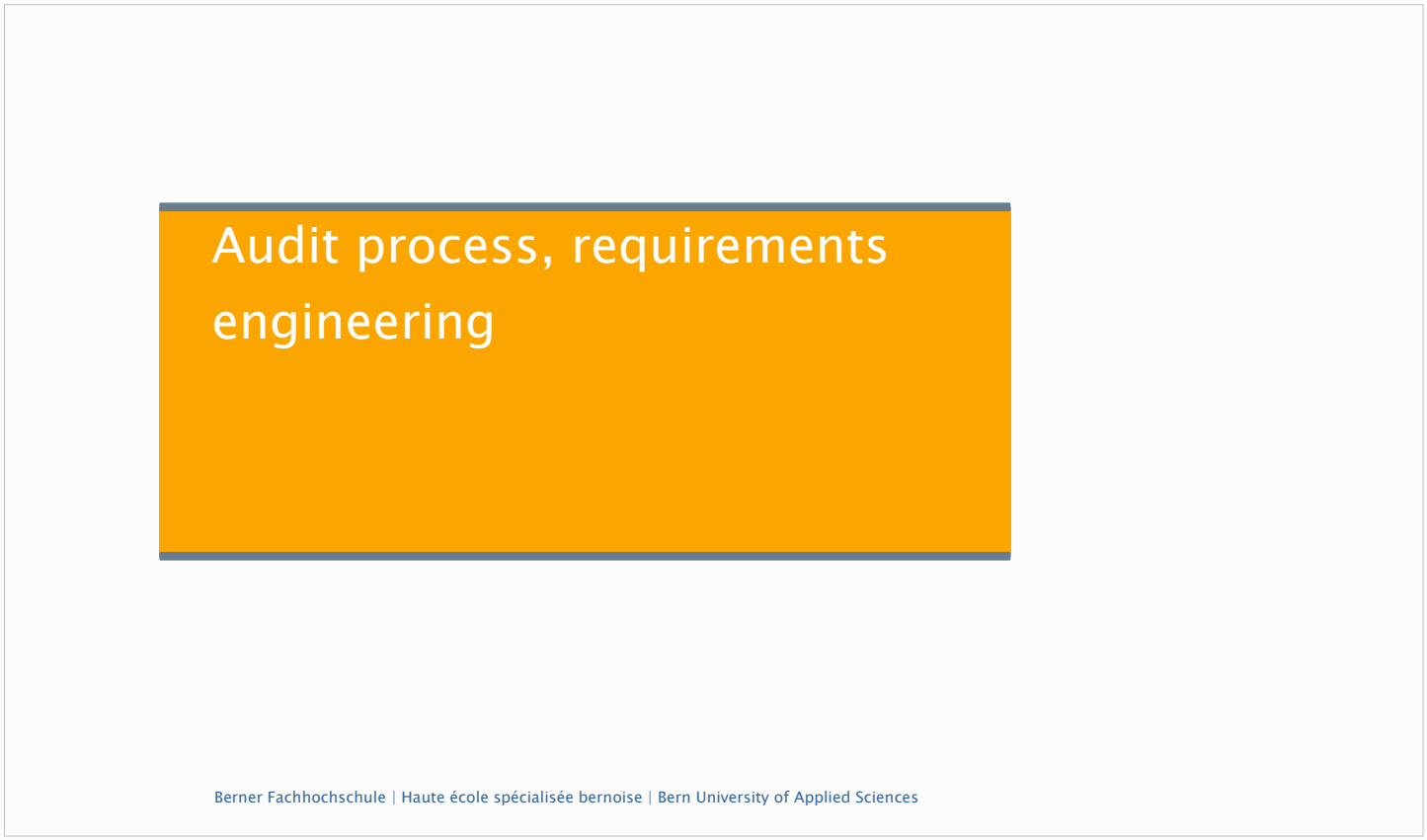
- Concern about the impacts of changing technologies, including support for new media and data formats
- Changing Designated Community

This period extends into the indefinite future.

Entsprechend auch die Definition, was Langzeit bedeutet:

Eine Zeit die genug lang ist, um sich mit dem technologischen Wandel beschäftigen zu müssen.

Und eine Zeit die lang genug ist, um sich mit einer verändernden Zielgruppe auseinander zu setzen.



Nach der Einarbeitung in die Grundlagen wurde das Public Domain Projekt einem Auditprozess unterzogen.

CCSDS 652.0-M-1 Audit

- The audit consists 108 metrics to test a digital repository for its ability for long time preservation and trustworthiness
- The audit was done to get a detailed view on the current status of the activities of the Public Domain Project

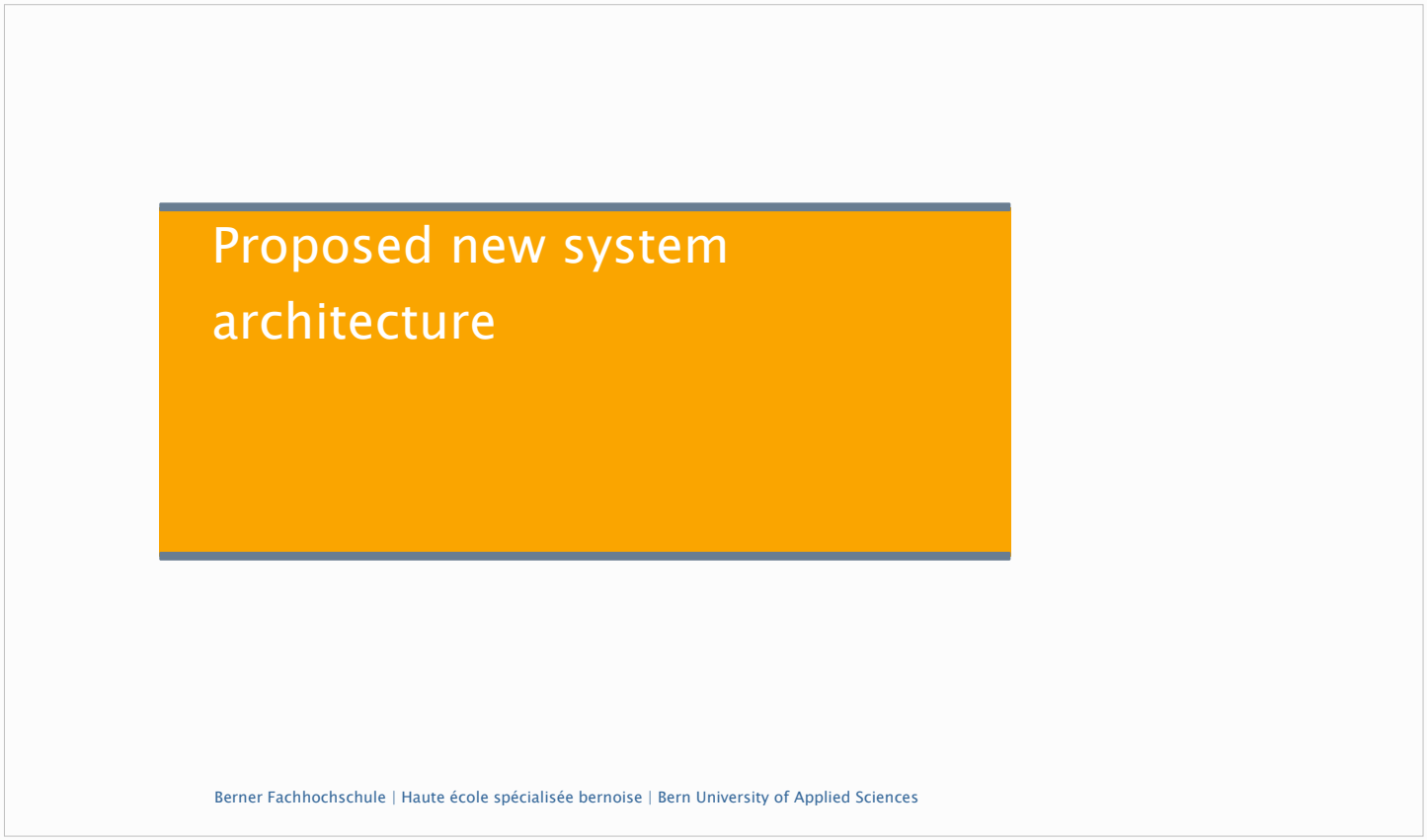
Damit konnte eine detaillierte Sicht auf den aktuellen Stand gewonnen werden.

Requirements engineering

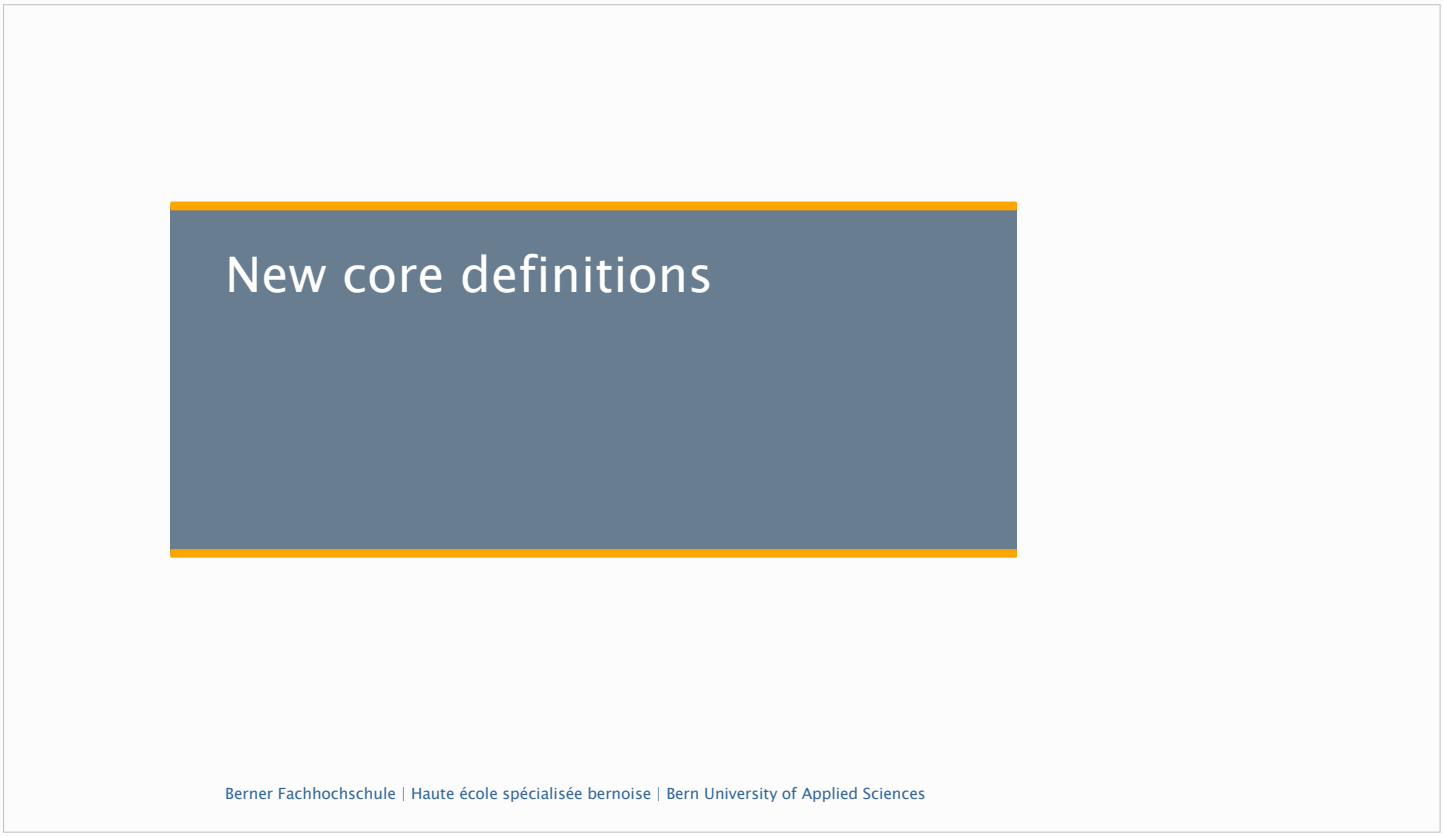
The specific requirements for a new system architecture for the Public Domain Project where defined based on:

- The definitions and requirements of the OAIS model
- The results of the audit

Die erarbeiteten Grundlagen und die Resultage des Audits wurden genutzt um die Anforderungen an eine Systemarchitektur zu definieren.



Nun komme ich zum eigentlichen Resultat dieser Arbeit, der vorgeschlagenen neuen Systemarchitektur.

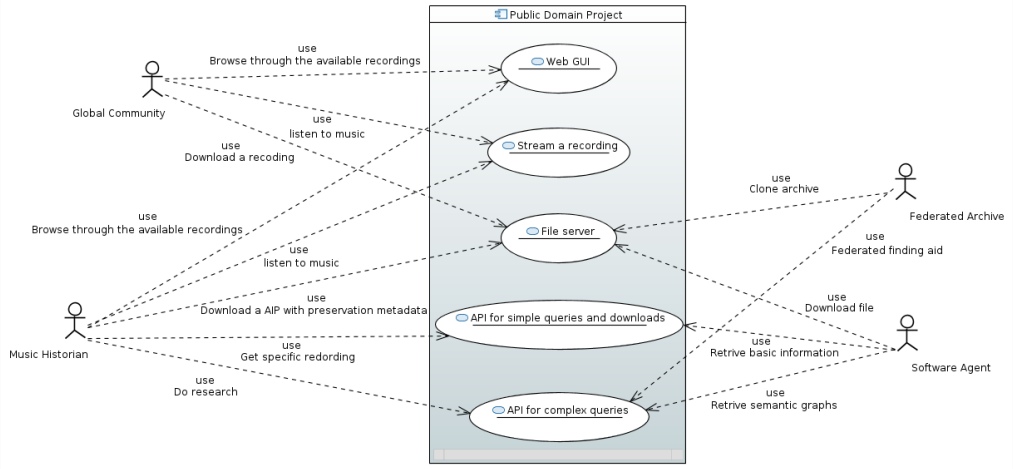


Ich kann hier nur kurz auf die drei wichtigsten Bereiche eingehen.

Die Kerndefinitionen beeinflussen alle Bereiche des Langzeitarchivs und mussten zu aller erst genau definiert werden.

New core definitions

Designated communities



Die Definition der Zielgruppen gibt die nötige Form und Tiefe der Metadaten vor und auch welche Art von Diensten erwartet werden.

New core definitions

Content information

The Public Domain Project takes on the responsibility to preserve the digital audio works that where transferred to it. The content information is defined as:

- The acoustic information in the frequency band that is audible by humans (15 Hz bis 20 kHz)
- All the needed metadata to determine the identity, provenience, origination and authenticity

Die Definition der Inhaltsinformation gibt das eigentliche Erhaltungsziel vor.

New archival information package (AIP)

Das neue Archivinformationspaket ist eine Weiterentwicklung des jetzigen Ansatzes.

Existing archival information package

- Flac file for audio data
  - Free lossless audio codec, Flac, is a open standard, open source and patent free
  - Well known format for producers and audiophile end users, uncommon in archives
- Wiki page for preservation metadata
  - Stored in database on separate computer
  - Not well suited for automated processing

Bisher besteht es aus einer Flac Datei und dazugehörigen Metadaten auf einer Wiki-Seite.

Proposed archival information package

- Matroska container (MKV) consisting of
  - Audio data in the Flac format
  - XML files for preservation metadata using
    - DublinCore, DCMI Abstract Model
    - PREMIS 3.0

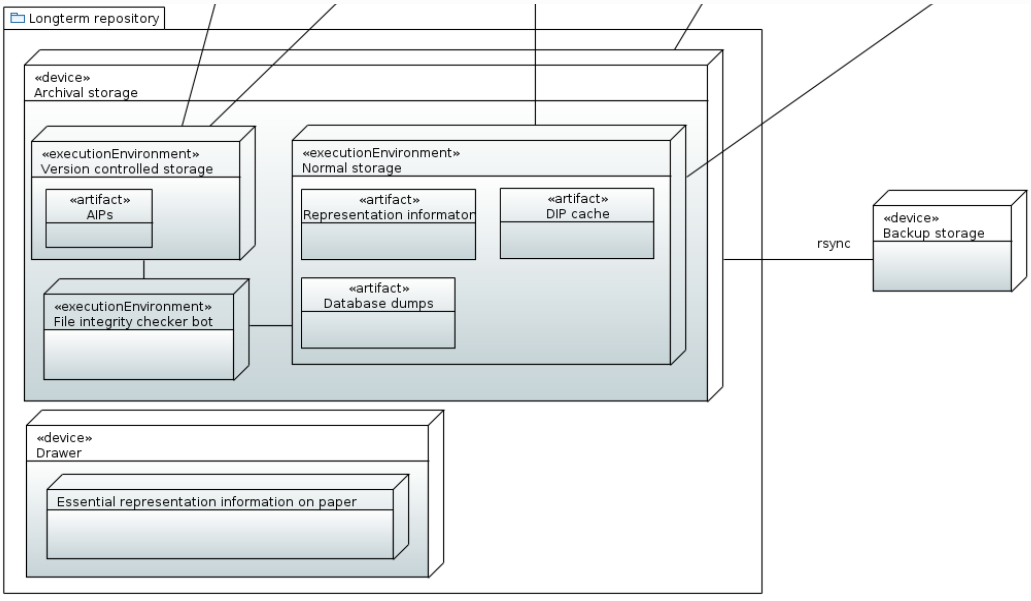
As the currently used Flac file, the MKV file can be directly played by usual audio players.

Neu soll all diese Information in einer einzelnen Matroska Datei untergebracht werden. Diese kann weiterhin in jedem üblichen Player abgespielt werden.

Archival storage system

Das Archivspeichersystem muss viele der vorher genannten Eigenschaften erhalten können.

Archival storage system



Dazu wird vorgeschlagen einen Speicherserver aufzubauen bei dem die Archivinformationspakete versionsverwaltet werden und alle Änderungen im Server und im Archivinformationspaket protokolliert werden.

Alle Daten werden regelmässig auf ihre Korrektheit überprüft.

Archival storage system

Preservation of representation information

- Source code based Gentoo GNU/Linux
  - All source code of the installed software is locally available
  - Representation information automatically stays in sync with the used software version
- Essential information on paper
  - Unicode table, standards used in AIP, SATA specification etc.

Um die Lesbarkeit und Darstellbarkeit zu gewährliesten wird der Server mit Gentoo Linux aufgesetzt. Bei dieser Distribution wird für alle Softwareteile der Souce Code lokal vorgehalten.

Zusätzlich wird neben dem Server eine Schublade installiert mit den essentiellsten Repräsentationsinformationen auf Papier.

Conclusion

Zusammenfassend kann gesagt werden...

Conclusion

A word on metadata standards

- Complex topic
  - Development from subject headings to linked open data and the semantic web
  - Several standards with $>$ 100 pages
- No consensous
  - In some areas accepted metastandards, but used with different vocabularies

(Wurde aus Zeitgründen übersprungen)

Conclusion

Results an outlook

- Gained a deep knowledge in the field
- The chosen approach and the used audit system fulfilled the expectations
- The master thesis is about implementing the proposed system architecture
  - A new audit will be done a the end of the master thesis to measure the progress

..., dass mit dieser Arbeit ein gutes Verständnis der digitalen Langzeiterhaltung aufgebaut werden konnte.

Der gewählte Ansatz mit dem Audit bracht den erwünschten ungetrübten Blick auf die aktuelle Situation.

Darum soll nach der Mater Thesis wieder ein Audit durchgeführt werden um den Fortschritt zu dokumentieren.

Thank you for your attention

For further questions you can contact me by e-mail:

nuess0r@pdproject.org

Berner Fachhochschule | Haute école spécialisée bernoise | Bern University of Applied Sciences

Ich danke für ihre Aufmerksamkeit und stehe nun für ihre Fragen zur Verfügung.

Audit conclusion

- 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

New core definitions

Designated communities

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, 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.

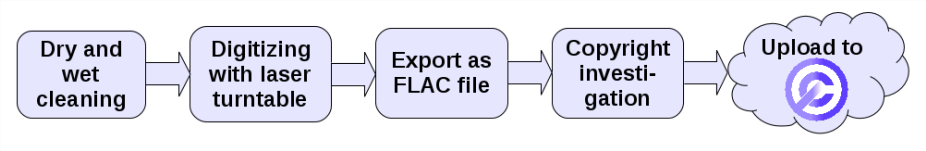
Proposed Archival information package

IETF cellar project

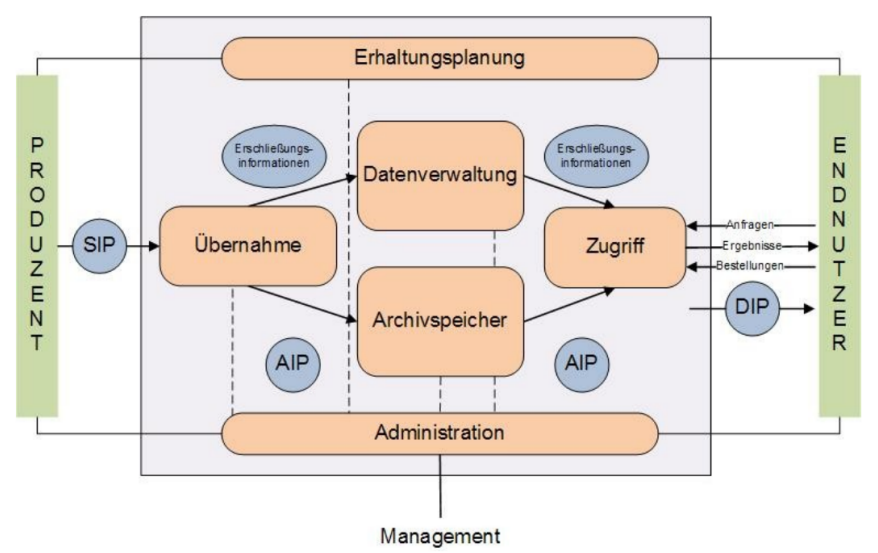
Using existing work done by the development communities of Matroska, FFV1, and FLAC, the Working Group will formalize specifications for these open and lossless formats. In order to provide authoritative, standardized specifications for users and developers, the Working Group will seek consensus throughout the process of refining and formalizing these standards

- https://datatracker.ietf.org/wg/cellar/charter/

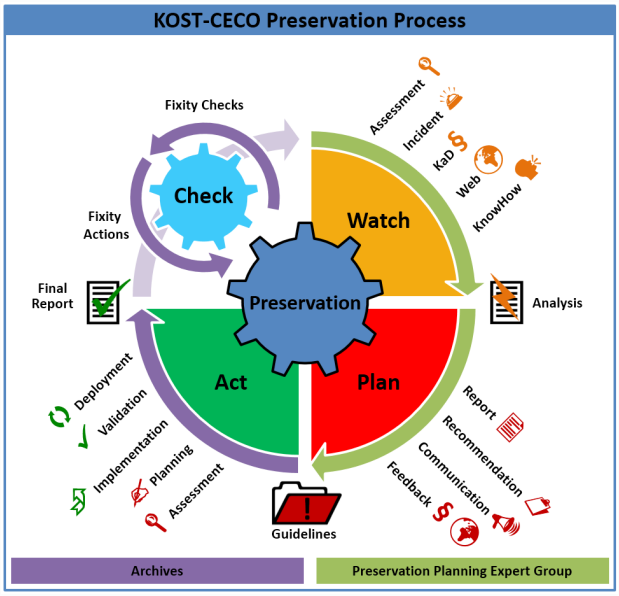
Digitization process



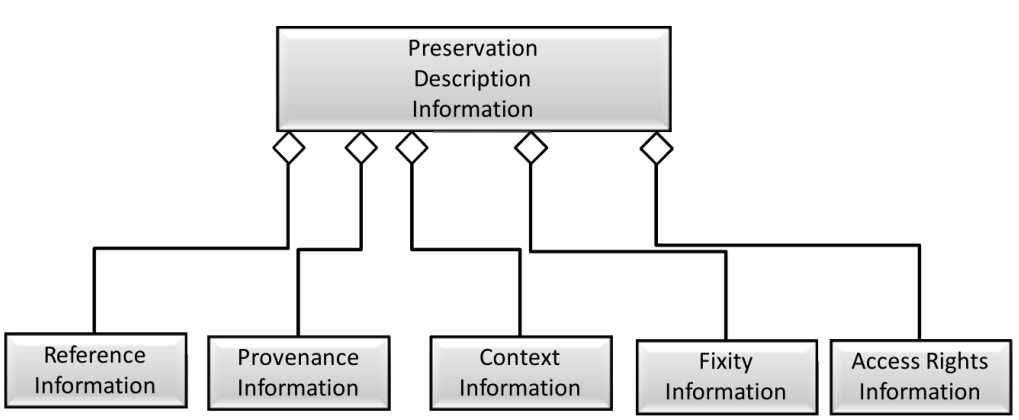
OAIS Functional Entities



Preservation process



Preservation Description Information



Supported by this experts in the field

- Christoph Müller, BSc in Information Science FHO
  - Consulting projects in the field of modern records management and digital long time safe keeping
- Hartwig Thomas, Dr. sc. math.
  - CEO Enter AG, Rüti