Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Long time archive for audio works

Project 2

- Institute for Human Centered Engineering HuCE

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

Introduction

Structure of this presentation

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

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

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

Basics of digital long time preservation

Paradigm shift in preservation

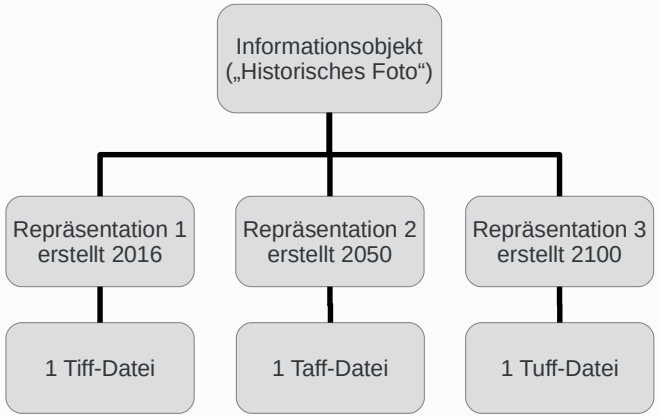
Preservation by analog media conservation

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

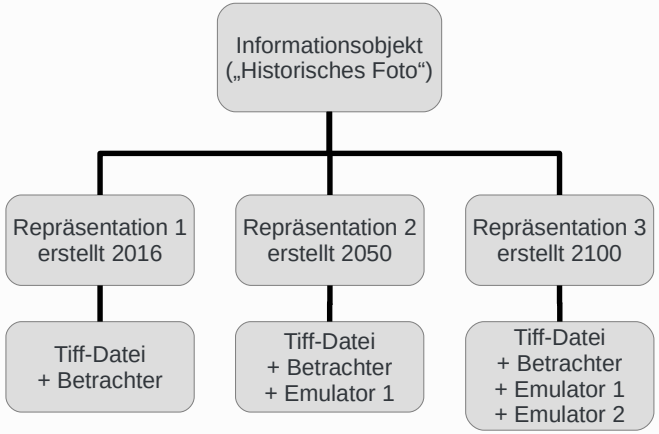
Paradigm shift in preservation

Preservation by digital migration

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



Transformation strategy



Emulation strategy

Simple example: *.txt file

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

Simple example: *.txt file

- An ASCII Table (Nowadays a Unicode table)

- Is that enough?

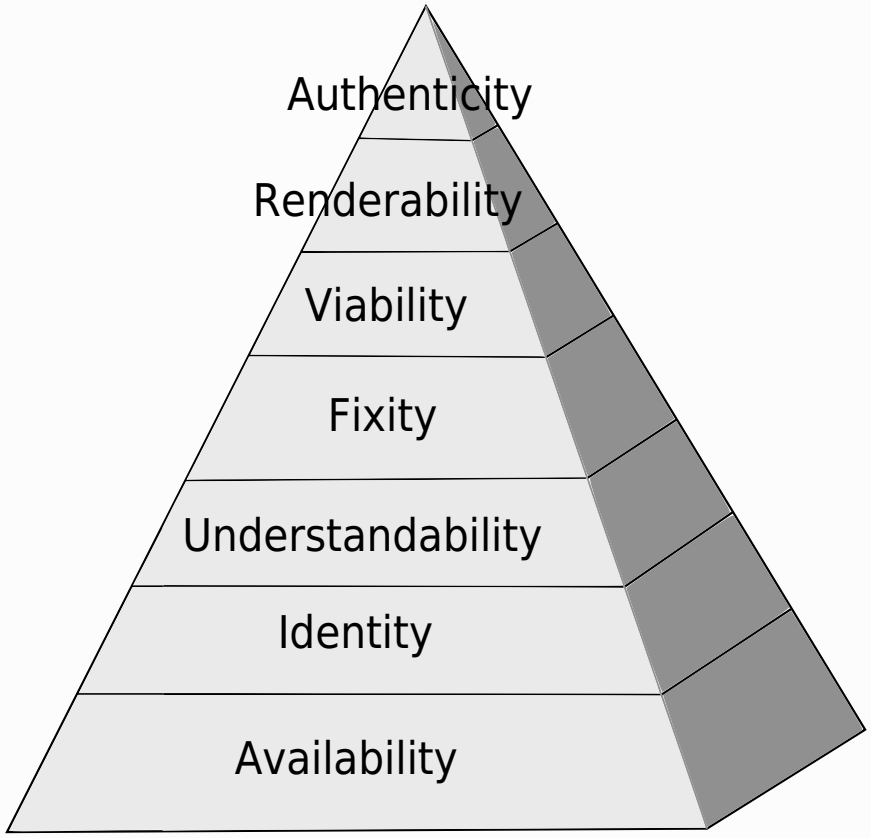
Simple example: *.txt file

No!

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

What needs to be preserved



Preservation Goals	**Means**
Authenticity	Authentication
Renderability	Format strategies
Viability	Media management
Fixity	Secure Storage
Understandability	Documentation
Identity	Description
Availability	Capture Selection

OAIS reference model

Reference Model For An Open Archival Information System (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.

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.

Audit process, requirements engineering

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

Requirements engineering

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

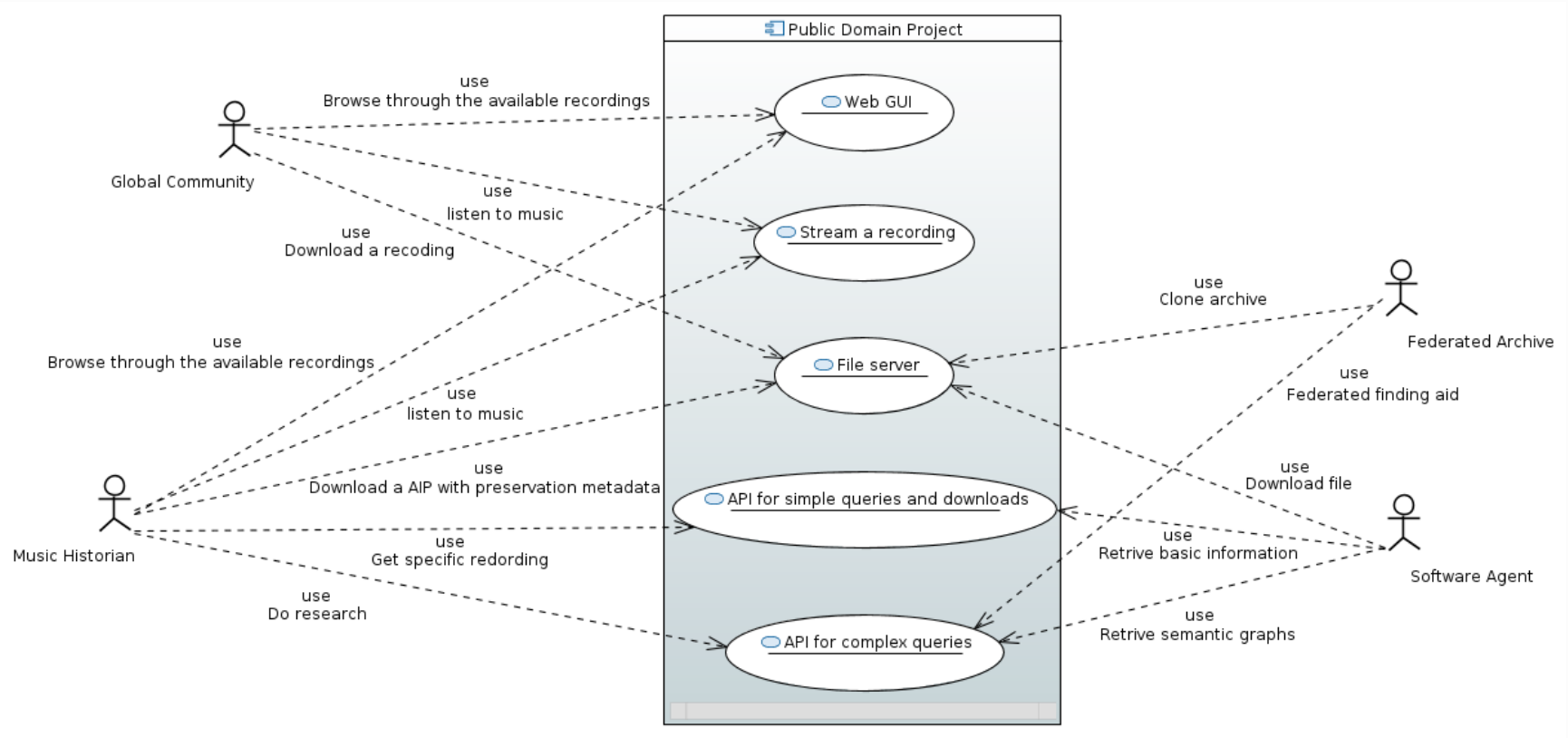
- The definitions and requirements of the OAIS model
- The results of the audit

Proposed new system architecture

New core definitions

New core definitions

Designated communities



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

New archival information package (AIP)

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

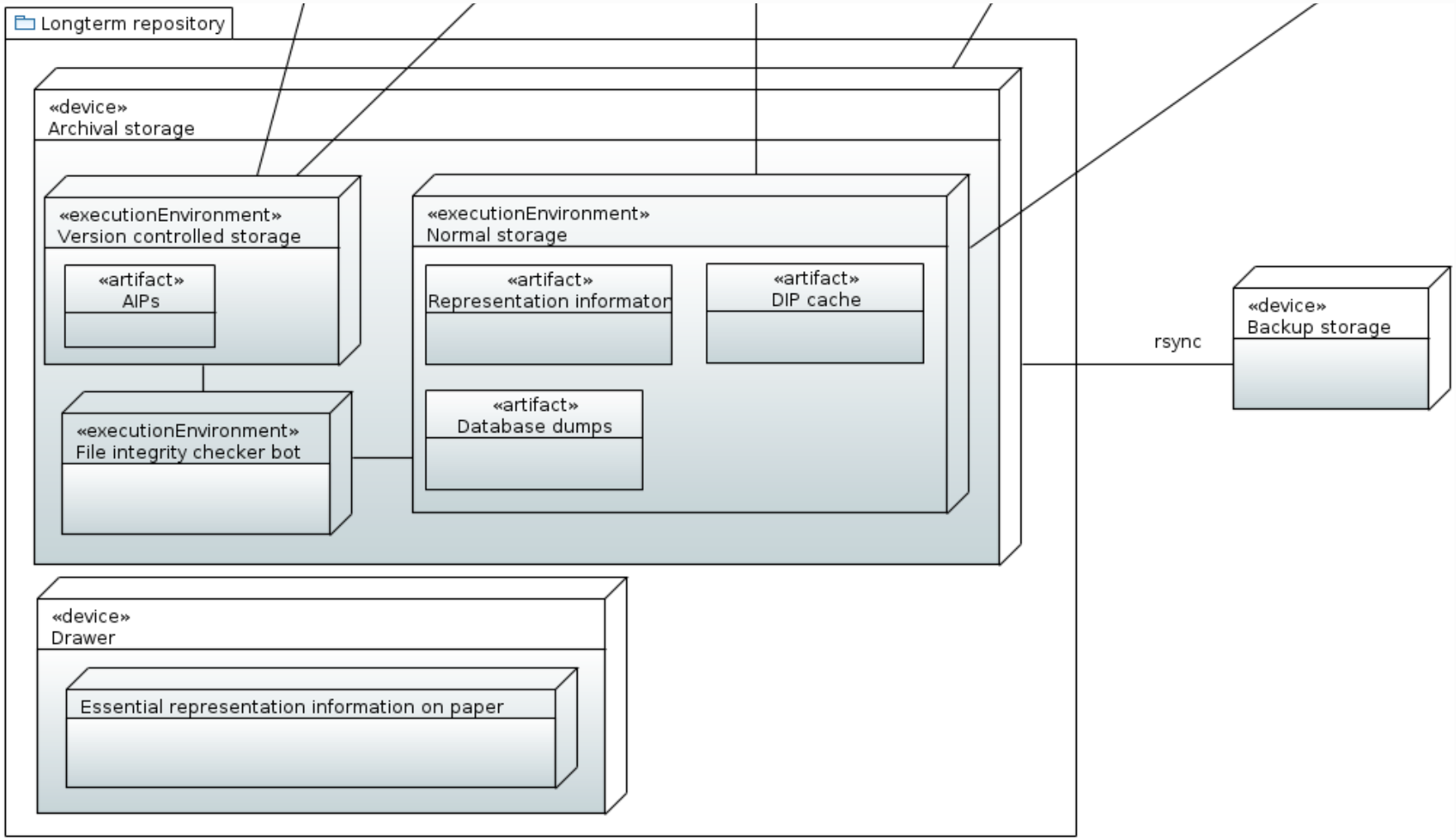
Proposed archival information package

- 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.

Archival storage system

Archival storage system



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.

Conclusion

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

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

Thank you for your attention

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

nuess0r@pdproject.org

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.

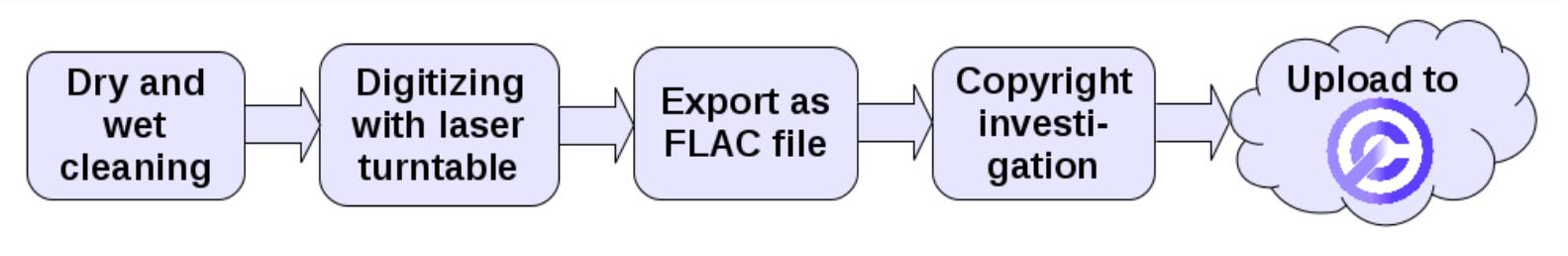
Proposed Archival information package

IETF cellar project

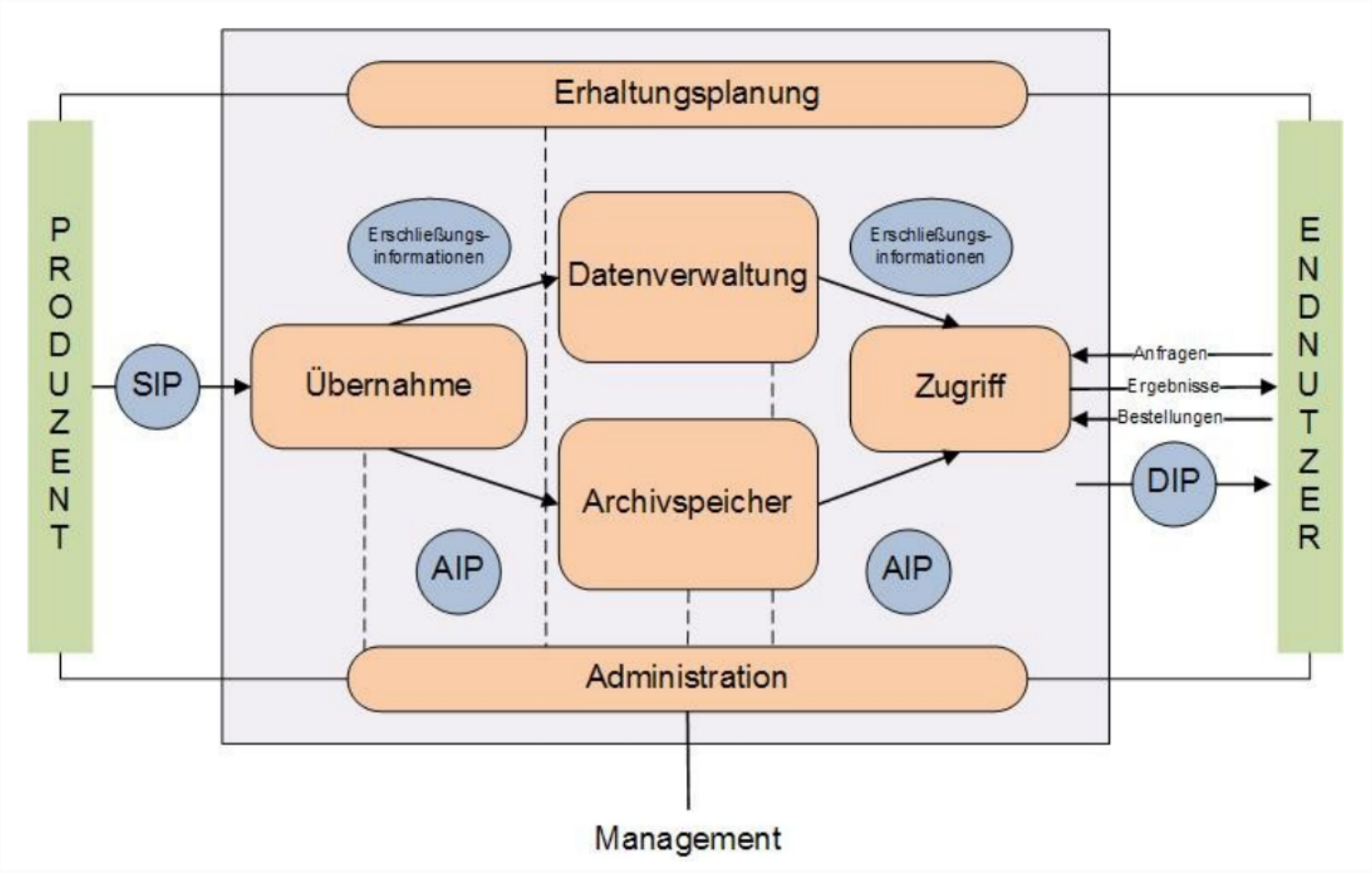
Using existing work done by the development communities of Matroska, FFV1, and FLAC, the Working Group will formalize specifications for these open and lossless formats. In order to provide authoritative, standardized specifications for users and developers, the Working Group will seek consensus throughout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

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

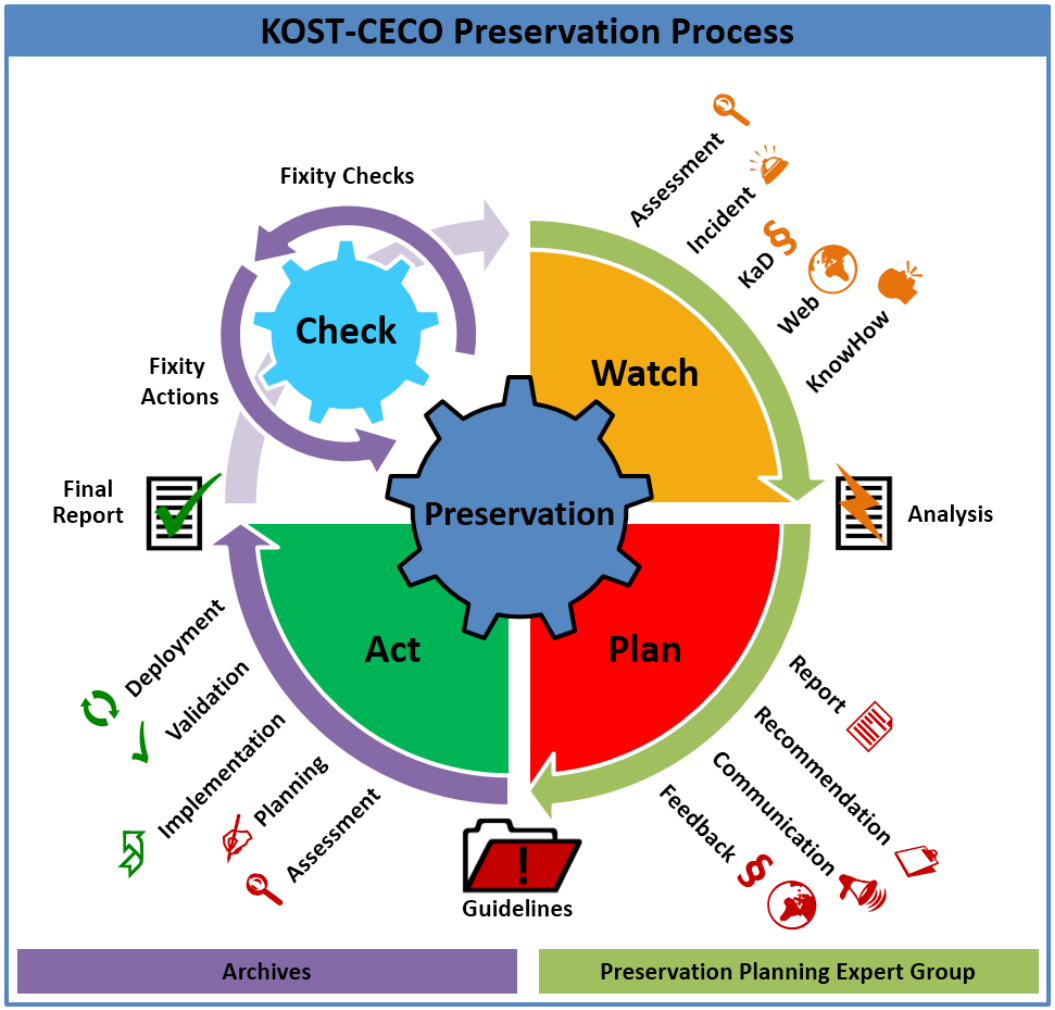
Digitization process



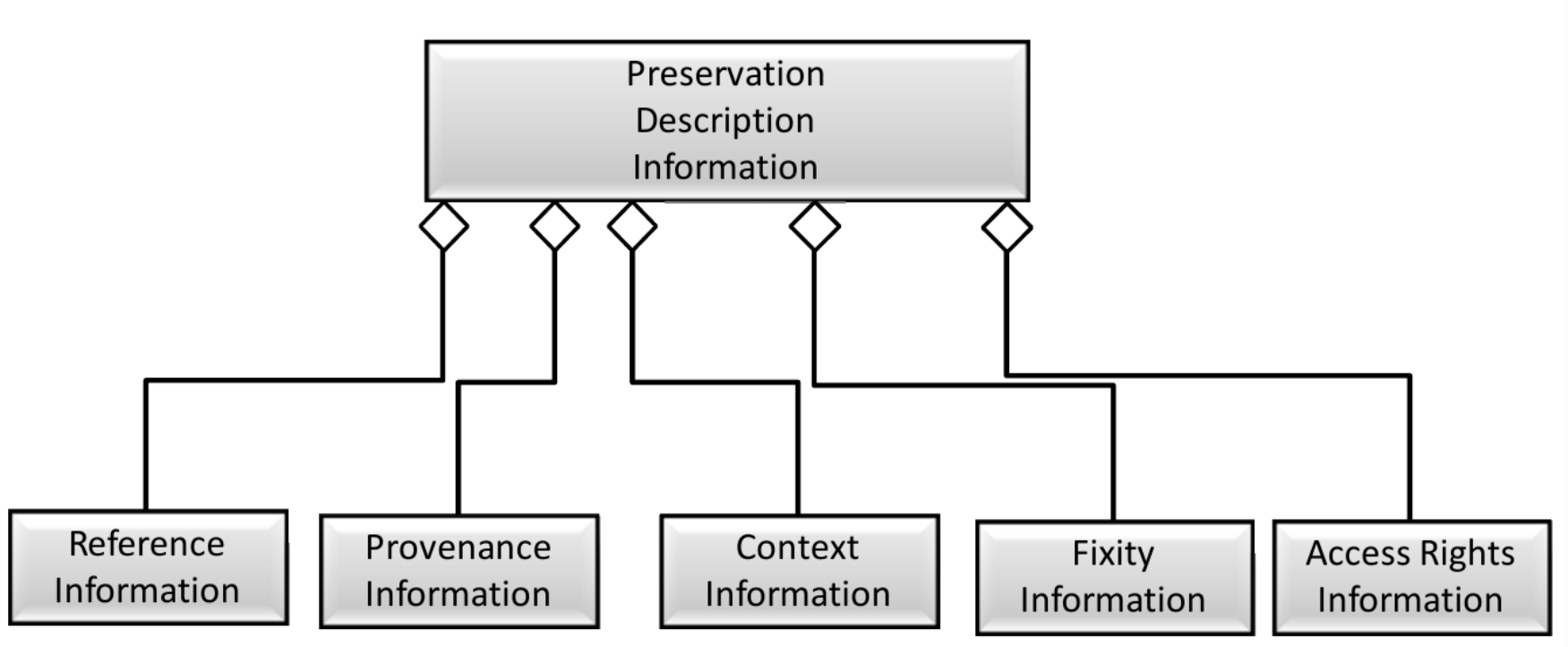
OAIS Functional Entities



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