

# Archival Ingest



# Scope of submission

Meaning: The stuff someone gives you to preserve (for them).

For example:

- 1-or-more content files
- Metadata
- Other (relevant) information

# What is a SIP?

*“An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.” — OAIS, 1-15*

# Shape of SLPs to come

- File
- Folder(-structure)
- ZIP, TAR, etc.
- BagIt Bag
- Digital objects + database entry
- A video tape, film, image, tape, ...
- ...anything! :D

# Mom... Where do SLPs come from?

- Creators
- Digitization vendors
- Staff
- Passionate collectors
- All of the above!
- and more...

# How to decide...

## what is accepted/required?

- Source (e.g., film, video)
- Agreements with producers
- Internal capabilities
- Metadata only available from producer
- Policies: collection, format
- etc...

# What is...?





# Ingest activities

## Speaker notes

These are as different as the world is colorful and amazing! ;)

They depend on things like:

- media type
- focus, capacities
- policies
- etc

- Which of these activities applies to which kind of data formats?
- Which steps could be “source preparation”? And for which kinds of materials?

# Typical ones

- Prepare & record (analogue) source
- Generate unique ID
- Validate or generate fixity/hash data
- Format policy checks
- Format normalization
- Create derivatives
- Create metadata
- Logging
- Virus check
- etc...

## Questions:

- For what reason(s) would you want/need an identifier?
- Why is it desired that it's unique?
- What is "Linked Open Data", and what's its relation to

# The “unique” Identifier

## A must-have !

Also known as:

- ID
- Object ID
- Item ID
- Archive signature
- **UID, UUID**
- ...

# The “unique” Identifier

## Examples

- V-00815
- W/S #00034
- FBW002984
- 38AF2EC1A13494B9DF6FD6E75960307
- 111-ADC-4319
- VHS-0317
- adBDwKf\_aSE
- Q83697636
- ...

### Speaker notes

All these identifiers are either from real situations, or based on them.

To calculate the number of different/unique combinations that an ID can depict, you can do this:

- How many “digits” can it have?
- How many different symbols does it offer?
- Then: symbols ^ digits = possibilities.

Remarks about IDs from Youtube and Wikidata:

- Youtube: This is an interesting one. By including upper/lowercase alphanumeric characters and some non-alphanumeric ones, it's actually possible to have a rather short, yet unique ID. It's *somewhat* human readable/handleable, but definitely that was not the main priority for that choice. Unique combinations at 11 digits with about 72 symbols:  $72^{11} = 269561249468963094528$
- Wikidata: I'm actually surprised that they chose a decimal-numerical choice only prefixed by a “Q”. It's kind of cool and simple, but will grow quite large in digits over time.

# Identifier: Considerations

- Distinguish which type of object/media?
- How many objects to expect (per time/year)?
- Human readable/handleable? (vs as unique as possible)
- Print on stickers on physical objects?
- Print as bar codes?
- How to “ingest” external collections into that schema?
- Does it scale *enough*?
- Valid for which duration?

# Format Normalization

- Popular SIP to AIP use case
- Improve preservation properties
- By switching to a “better” format
- Cleaning/normalizing data (dialects)

# Format Normalization

## Examples

- Rewrap container (eg MKV, MOV, MXF)
- Audio to PCM
- Convert video to FFV1, V210, H.264, etc

In order to decide which files can be ingested without format normalization, or need other special attention, it is good to define so called rules which properties a media file shall have. These rules are (also) called “policies”.

These policies can then be used to decide further steps to be taken for certain objects.

# Format Policy Checks

- Define conditions for tech-MD properties
- “whitelist” formats
- Spot irregularities



# MediaConch

MediaConch

CheckerPoliciesPublic PoliciesDisplayDatabaseSettingsHelp

## Check files

Check local file

Check online file

Check local folder

Policy

No policy

Display

MediaConchHtml

Verbosity

5

☐ Enable fixer ⓘ

☐ Full parse ⓘ

Select files

Choose Files

No file selected

Check files

Files added successfully

## Results

Apply a policy to all results

Choose a new policy to apply

Show

10

entries

Search:

Files	Implem	Policy	MediaInfo	MediaTrace	Status
1_one.avi	✓ Valid	✓ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analyzed
2_two.avi	✓ Valid	✓ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analyzed
3_three.avi	✓ Valid	✗ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analyzed
4_four_A.avi	✓ Valid	✗ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analyzed
4_four_B.avi	✓ Valid	✗ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analyzed
5_five.avi	✓ Valid	✗ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analyzed
6_six_A.avi	✓ Valid	✓ PAL SD	👁️ ⚙️	👁️ ⚙️	✓ Analvzed

Speaker notes

No notes on this slide.

# MediaConch

MediaConch

CheckerPoliciesPublic PoliciesDisplayDatabaseSettingsHelp

Policy editor

Policy list:

Search

User policies

Austrian Mediathek: Preservation Master (Video) (and)

Audio: Good quality resolution (and)

Audio properties check (and)

PAL SD (and)

Resolution is SD (and)

Width is 720px

Height is 576px

Has 25 fps

MPEG-4 AVC (H.264/AAC) (and)

System policies

Is this NTSC or PAL SD? (and)

Example MKV FFV1 digitization policy (and)

Matroska is well described? (and)

CAVPP Preservation Master (and)

Memoriav Video files Recommendations (or)

TN2162 compliant? (and)

Edition:

Policy name ⓘ

PAL SD

Policy description ⓘ

Policy type ⓘ

AND

Policy license ⓘ

Creative Commons Attribution-ShareAlike

Save policy info

Add a new policy

Add a new rule

Actions:

Duplicate

Export

Delete

Add this policy to MediaConchOnline ⓘ

Add as public

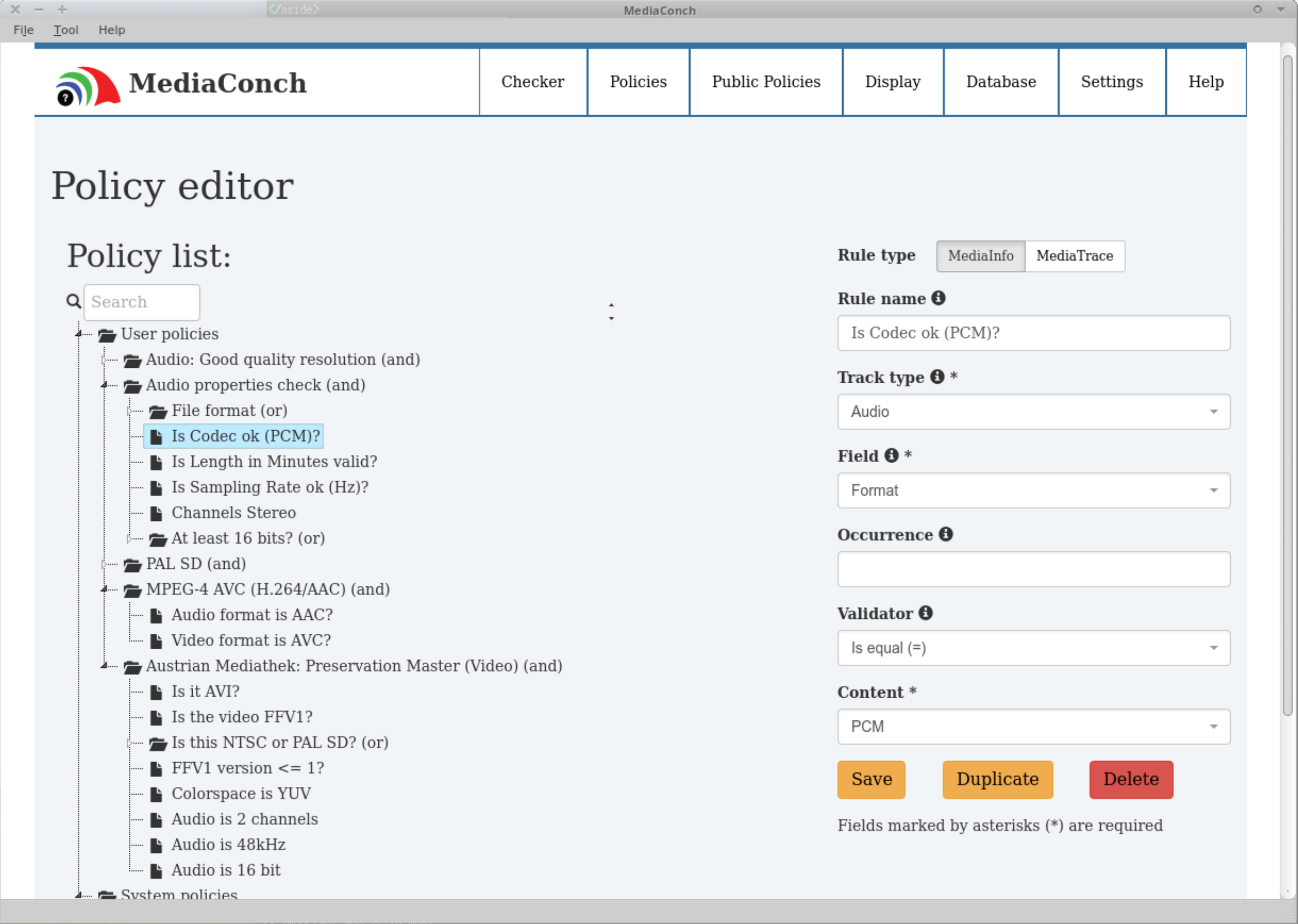
Add as private

Speaker notes

No notes on this slide.

18

# MediaConch



Speaker notes

No notes on this slide.

# Possible Bottleneck

*“Ingest can be a dangerous bottleneck.  
Don’t let the perfect be the enemy of the  
good”*

**Comments?**  
**Questions?**