Topic 4 - Ingest

Speaker notes

No notes on this slide.

Speaker notes

No notes on this slide.

Scope of submission

Speaker notes

No notes on this slide.

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

Speaker notes

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

For example:

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

Shape of SIPs to come

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

Speaker notes

No notes on this slide.

Mom... Where do SIPs come from?

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

Speaker notes

No notes on this slide.

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…

Speaker notes

Source: Example: If it’s a physical medium “film” it goes to a film archive, if it’s a magnetic tape it may go somewhere else: Mediathek.

This may also be a(n unwritten) policy.

What is...?

Speaker notes

No notes on this slide.

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

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…

Speaker notes

Questions:

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

The “unique” Identifer

A must-have !

Also known as:

- ID
- Object ID
- Item ID
- Archive signature
- UID, UUID
- …

Speaker notes

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” Identifer

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?

Speaker notes

No notes on this slide.

Format Normalization

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

Speaker notes

No notes on this slide.

Format Normalization

Examples

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

Speaker notes

No notes on this slide.

Format Policy Checks

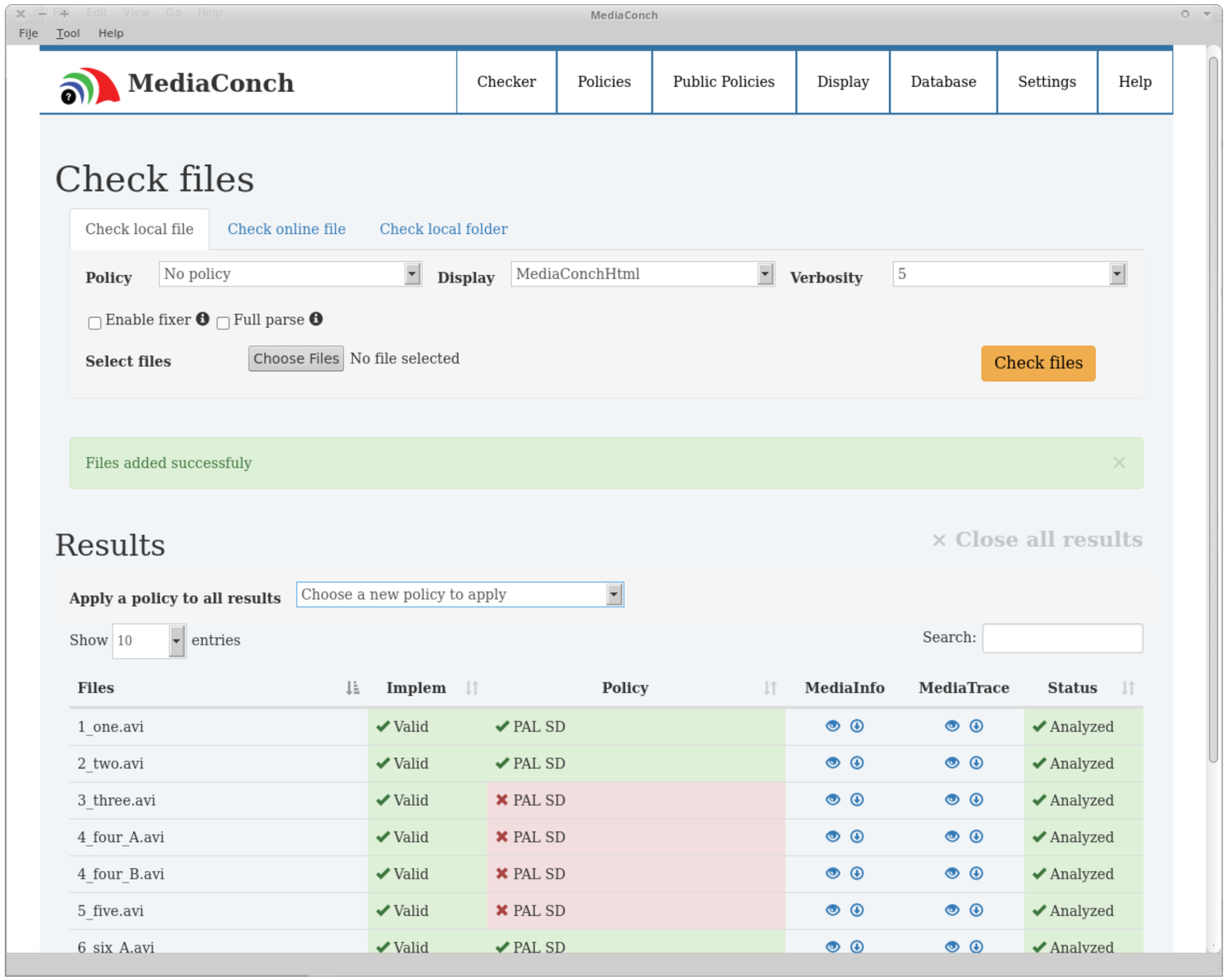
- Define conditions for tech-MD properties
- “whitelist” formats
- Spot irregularities

Speaker notes

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.

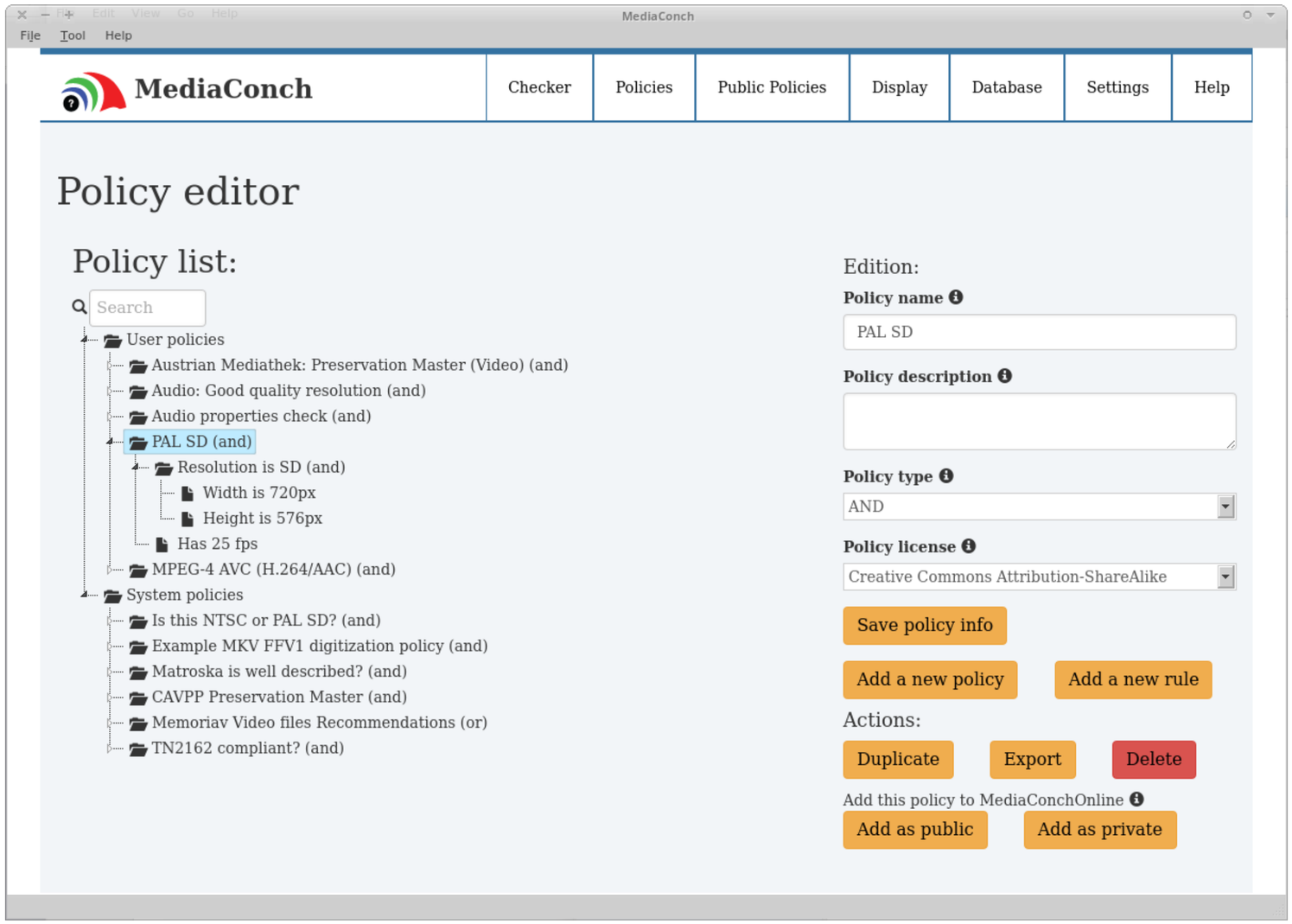
MediaConch



Speaker notes

No notes on this slide.

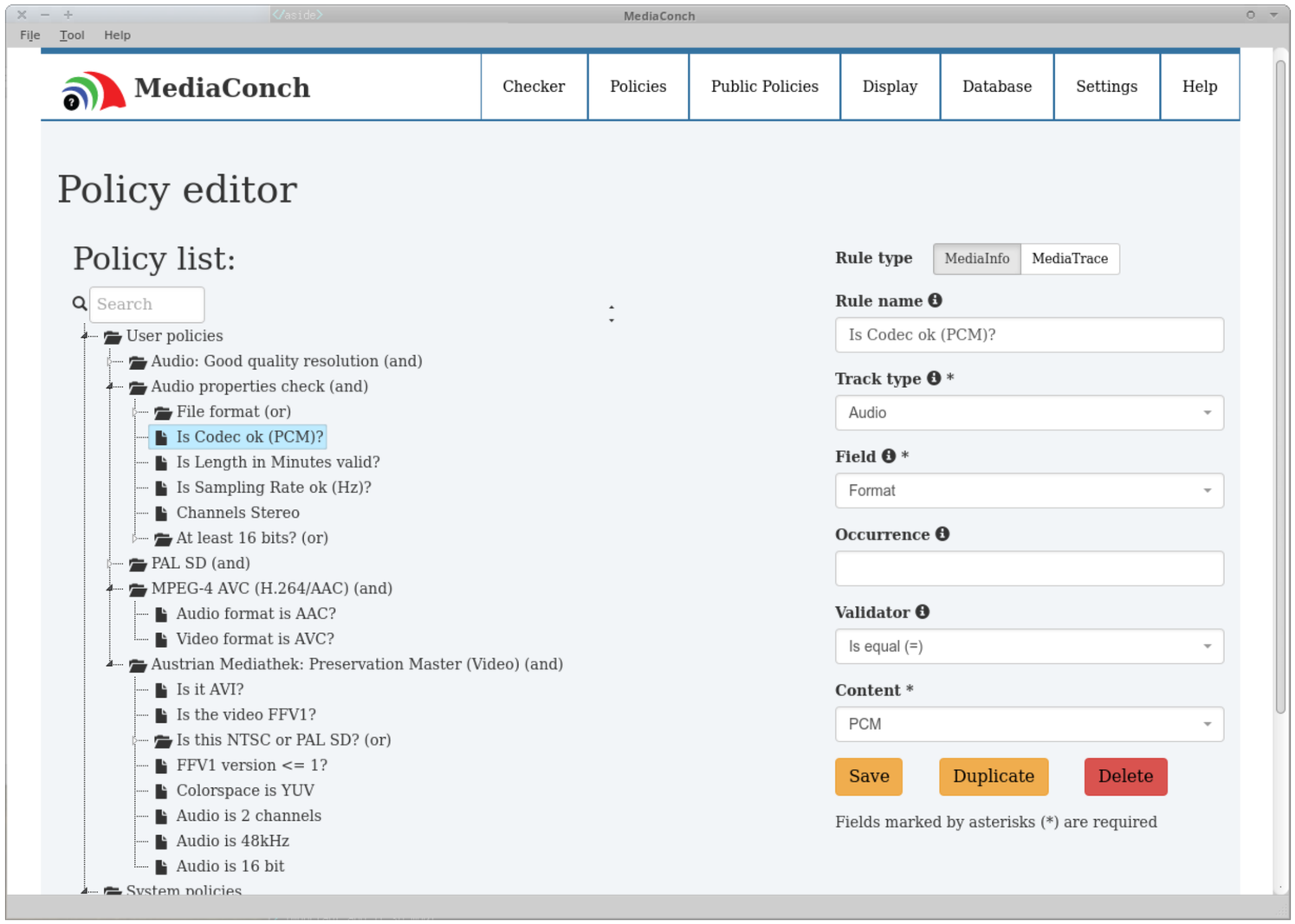
MediaConch



Speaker notes

No notes on this slide.

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”

Speaker notes

No notes on this slide.

Comments?

Questions?

Speaker notes

No notes on this slide.