SFA SIP Validation Enhancements

Proposed error message updates - REVISED based on SFA feedback, 2024-08-14

Overview

The original SIP validation in Enduro was based on the Python scripts originally developed by Christa. ASI have rewritten these in Golang (a different programming language), but as for the validation checks performed and the error messages returned, these have so far been largely unchanged.

Our initial Talk Aloud testing in Sprint 2 showed a lot of confusion over what to do when encountering the current error messages. On 2024-07-18, Dan drafted up some proposed revisions to the current validation error messages based on initial SFA input. SFA then had two different groups review these proposals and provide additional feedback (here and here in Drive). Below is a summary of the proposed enhancements, to be delivered for SFA review in a future sprint.

Common change requests to all error messages

- **Use relative paths instead of absolute ones**, starting from the package itself
  - Talk-aloud testing showed that users tended to stop reading error messages when encountering long absolute paths, seeing it as too technical
  - This often meant missing information on the exact file with issues, or suggested next steps
  - Relative paths from the package should shorten this and keep the focus on the package structure itself, removing extraneous information about the server directory structure

- **Break up the error message visually with linebreaks,** for easier readability
  - Add a bit more space between the high-level error message, the details, and the final path
  - **If possible, style output a bit:**
    - Bold the high-level message
    - Use escaped courier text for file names and paths

- **Add a general suggestion on next steps**
  - Users wanted guidance on what to do next, now that a validation failure has been encountered
  - We should try to include some general suggestions on follow-up actions
  - E.g. "Please review the file and ensure that it matches eCH-0160 requirements" or similar

- **Separate the file path from the error message**
  - Keep the error message itself as readable as possible
  - Add the relative path below as additional information

- If possible, **include a link to the relevant package in the failed directory**

- ○ At minimum, tell users that the package has been moved to the failed-sips bucket in minIO
  - ■ Make this a hyperlink to the bucket itself
- ○ **Better yet:** if possible, provide a download link for the package directly, so the operator does not need to go to minIO first and find the relevant SIP

- ● **Number and group errors if multiple are returned in one check**
  - ○ For example, the file format messages - see the examples below
  - ○ Whenever a single validation error message includes more than one specific error, be sure to enumerate them clearly, matching the numbers of the error message body to the numbering used below for the filepaths

Validate metadata

Current error message example:

- ● Content error: metadata validation has failed: /var/lib/enduro/preprocessing/enduro2684964344/extract107201249/package-3/additional/UpdatedAreldaMetadata.xml does not match expected metadata requirements

---

Specific recommendations

- ● **Include the line number** from the relevant metadata file if possible

Updated example proposal

Content error: metadata validation has failed.

`UpdatedAreldaMetadata.xml` does not match expected metadata requirements. Please review the file and ensure that it matches eCH-0160 requirements.

Error Location:
- ● `package-3/additional/UpdatedAreldaMetadata.xml`
- ● Error encountered on: `line 43`

Your package has been moved to the failed-sips directory. You can download it for review.

Validate file formats

Current error message example:

- Content error: file format validation has failed. One or more file formats are not allowed: file format "fmt/43" not allowed: "/var/lib/enduro/preprocessing/enduro2250323578/extract1009993925/little_vecteur_sip_failure_v2-extrafiles/content/d_0000001/attack-rabbit.jpg"

Specific recommendations:

- **Show file format name in addition to PRONOM ID**
    - Previously, only the PRONOM ID was included in the error message - while this is more precise most archivists do not have PRONOM IDs memorized. It would be helpful to also include the format name, which Sigfried provides.

- **Number outputs**
    - This particular check can return more than one error at a time. To ensure that users can properly align the separated relative path of the file from the main error message, we number both lists

Updated example proposal

Content error: file format validation has failed

One or more file formats found in the SIP are not allowed:

1. JPEG File Interchange Format - `fmt/43`
2. Graphics Interchange Format - `fmt/4`

Please review the contents of the SIP and ensure that only approved formats are included.

Location(s):

1. `package-3/content/d_0000001/attack-rabbit.jpg`
2. `package-3/content/d_0000001/some-other-file.gif`

Your package has been moved to the failed-sips directory. You can [download it](#) for review.

Validate structure

Current error message examples (4 different possible errors):

- Content error: SIP structure validation has failed: XSD folder is missing

- Content error: SIP structure validation has failed: Unexpected directory: "/var/lib/enduro/preprocessing/enduro2563181936/extract3119941294/package-3/some-other-thing"

- Content error: SIP structure validation has failed: Content folder is missing

- Content error: SIP structure validation has failed: metadata.xml is missing

Specific recommendations:

- **If possible provide the expected location for missing directories**
  - ○ Currently the messages just say what file is missing. If possible, indicate where the validation scripts are expecting to find these directories

Updated example proposals

Content error: SIP structure validation has failed

XSD folder is missing. Please review the SIP and ensure that its structure matches eCH-0160 specifications.

Expected location: `package-3/header/xsd`

Your package has been moved to the failed-sips directory. You can download it for review.

Content error: SIP structure validation has failed

Content folder is missing. Please review the SIP and ensure that its structure matches eCH-0160 specifications.

Expected location: `package-3/content`

Your package has been moved to the failed-sips directory. You can download it for review.

Content error: SIP structure validation has failed

Unexpected directory found. Please review the SIP and remove any extra directories to ensure that its structure matches eCH-0160 specifications.

Location: `package-3/some-other-thing`

Your package has been moved to the failed-sips directory. You can download it for review.

Additional work

NOTE: per issue [SFA-35](SFA-35), we are also adding additional checks to ensure that every checksum listed in the metadata file has a corresponding object with a matching checksum, and that there are no objects included that are not listed in the metadata file.

It would be good to see how these new error messages are currently structured, and update them to match the proposals above.

Technically we could ALSO be checking the directory structure against the structmap outlined in the metadata file - I don't believe that the current "unexpected directory" check is actually doing this. However, there is no issue filed for this work at present. Note that issue SFA-32 could in part be solved by adding this check.

In any case - if we have additional error messages, then the Common change requests listed above should be applied, and for any specifics, consult the team for discussion.