

---

## JHOVE's PDF-hul Module

- **Product:** Rosetta
  - **Product Version:** 5.3
  - **Relevant for Installation Type:** Local
- 

### Question

How is JHOVE's PDF-hul Module utilized by Rosetta?

### Answer

PDF files tested directly in JHOVE (v1.9, v1.11, and v1.16) often present no errors.

Rosetta v5.3 includes JHOVE v1.10 with the v1.7 PDF-hul Module.

Ex Libris Development confirms that JHOVE 1.16 can be used with the metadata extraction plugins in Rosetta.

However, in some cases technical metadata can't be extracted from some PDF files.

You may see the following error messages in TA Workbench Validation (e.g.):

1. Invalid object number in cross-reference stream,Failed to retrieve extractor properties.
2. Expected dictionary for font entry in page resource.
3. Improperly constructed page tree,Annotation object is not a dictionary.

To address these errors in the short-term, please create a "Format Identification Correction" rule to ignore it.

In the Management module navigate to Home > Submissions > Rules > Format Identification Correction to establish the rule.

### Additional Information

With the exception of the 1.6, 1.9 and 1.11 framework releases, every JHOVE release has seen updates to the PDF-hul Module.

JHOVE's most recent v1.16 included PDF-hul v1.8, which fixed two major bugs.

These lead to false validation errors relating to invalid page dictionary objects and improperly constructed page trees.

While a number of fixes have improved PDF/A validation, JHOVE has been proven unsuitable for PDF/A validation.

The coverage of PDF versions hasn't changed since PDF-hul 1.0; for "plain" PDF, JHOVE supports PDF 1.0-1.6.

Note that the metadata extraction plugins are part of the Format Library.

To formalize a more long-term solution, please consult directly with the Rosetta Format Library Working Group (FLWG) in order to update the Format Library not to report these issues.

---

- **Article last edited:** 15-Dec-2017