JHOVE2:
Taking technical metadata further

Hannah Frost
JHOVE2 Project Team
California Digital Library, Portico, Stanford University
The preservation challenge

Managing the gap between what you were given and what you need

- That gap is only manageable if it is quantifiable
- Characterization tells you what you have, as a stable starting point for iterative preservation planning and action

The preservation challenge

< 1/3 of respondents in the 2009 Planets survey felt they had control over the digital content they were being asked to manage

Planets, Survey Analysis Report, IST-2006-033789, DT11-D1, 2009-05-06
www.planets-project.eu/market-survey/reports/

– How do you know what you have?
– How can you verify that you received what you expected?
– How can you classify for analysis, planning, and workflow?
“Tell me about yourself...”

© United Features Syndicate, Inc.
JHOVE2 is ...

... a project to develop a next-generation open source framework and application for format-aware characterization

... a collaborative undertaking of the California Digital Library (CDL), Portico, and Stanford University

... a two year grant from the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP).
“What? So what?”

Characterization is the automated determination of the intrinsic and extrinsic properties of a formatted object

- Identification
- Feature extraction
- Validation
- Assessment

What? So what?
Validation vs. assessment

Validation is the determination of the level of *conformance* to the normative requirements of a format’s authoritative specification

– To the extent that there is community consensus on these requirements, validation is an *objective* determination

Assessment is the determination of the level of *acceptability* for a specific purpose on the basis of locally-defined policy rules

– Since these rules are locally configurable, assessment is a *subjective* determination
“We report, you decide...”
Characterization in ingest workflows
JHOVE2 feature set

Multi-stage processing

- Signature-based identification
  ✓ DROID
  http://droid.sourceforge.net/
- Feature extraction
- Validation ✓
- Message digesting
  ✓ Adler-32, CRC-32, MD2, MD5, SHA-1, SHA-256, SHA-384, SHA-512
- Rules-based assessment
**JHOVE2 feature set**

Processing of objects spanning files and objects that are subsets of files

Recursive processing of objects arbitrarily-nested within containers

Extensive configurability

Complete documentation

- User’s guide
- Programmer’s guide
- Architectural overview
- Module specifications
Supported formats

JHOVE2 can identify (by DROID) many more formats than it can validate (by modules)

– PRONOM registry documents over 550 “formats”

http://www.nationalarchives.gov.uk/PRONOM
Supported formats

ICC color profile
JPEG 2000
PDF
SGML
Shapefile
TIFF
UTF-8
WAVE
XML
Zip

- ICC color profile
  - (ICC.1:2004-10)
- JPEG 2000
- PDF
  - PDF 1.0 – 1.7, ISO 3200-1, PDF/A-1 (ISO 19005-1), PDF/X-1 (ISO 15920-1), -1a (ISO 15930-4), -2 (ISO 15930-5) -3 (ISO 15930-6)
- SGML
- Shapefile
  - Main, Index, dBASE, ...
- TIFF
- UTF-8
  - ASCII (ANSI X3.4)
- WAVE
  - BWF (EBU N22-1997)
- XML
- Zip
Sample output (abridged)

<j2:feature name="FormType_raw" fid="http://jhove2.org/terms/property/org/jhove2/module/format/riff/RIFFChunk/FormType_raw"
 fidns="JHOVE2">
   <j2:value>WAVE</j2:value>
</j2:feature>

<j2:feature name="FormType_descriptive">
   <j2:value>Waveform Audio Format</j2:value>
</j2:feature>

<j2:feature name="Identifier">
   <j2:value>RIFF</j2:value>
</j2:feature>
Sample output (abridged)

<j2:feature name="Size">  
  <j2:value>1795916</j2:value>  
</j2:feature>  
<j2:feature name="FormatCategory_descriptive">  
  <j2:value>Microsoft Pulse Code Modulation (PCM) format</j2:value>  
</j2:feature>  
<j2:feature name="NumChannels">  
  <j2:value>2</j2:value>  
</j2:feature>
Sample output (abridged)

<j2:feature name="SamplingRate" funit="samples/second">
  <j2:value>48000</j2:value>
</j2:feature>

<j2:feature name="AverageBytesPerSecond">
  <j2:value>288000</j2:value>
</j2:feature>

<j2:feature name="BitsPerSample">
  <j2:value>24</j2:value>
</j2:feature>

<j2:feature name="isValid">
  <j2:value>true</j2:value>
</j2:feature>
Assessment

Evaluation of prior characterization information relative to local policy

Assessment results can inform preservation decision making

– Determine level of risk
– Assign level of service
– Take action now or later
Assessment rules

Assertions whose terms are logical expressions based on prior characterization properties

- Presence/absence of a property
- Constraints on property values
- Combinations of properties/values

The evaluation of the assertion results in new characterization properties

- Custom metadata that has significance in a local context
Assessment implementation

Each format module has a default rule set

Rules are logical expressions of the form:

If condition then consequent else alternative

Rules are configured using arules

– Utility developed by CDL to create rule set in XML
– Future plans: a GUI
Assessment

JPEG 2000 rule example (*pseudo-code*)

If ALL_OF
    validity == true;
    exists(colourBox);
    exists(resolutionBox.capture)
Then
    Acceptable
Else
    Not acceptable
End If
Assessment

TIFF rule example

If ANY_OF
   validity == true ;
   ((ifh.messages contains 
      'offsetNotByteAligned') or
    (ifd.messages contains 
      'offsetNotByteAligned') or
    (ifd.messages contains 
      'dateNotWellFormed'))
Then
   Acceptable
Else
   Not acceptable
End If
**Assessment**

WAVE rule example

```plaintext
If ALL_OF
    validity == true;
    exists(broadcastWaveExtensionChunk);
    waveFormatChunk.nSamplesPerSec == 96000;
    waveFormatChunk.nBitsPerSample == 24
Then
    Acceptable
Else
    Not acceptable
End If
```
Assessment

XML rule example

If ANY_OF
  validity == true ;
  (validity == undetermined) and
  (wellFormed == true)
Then
  Acceptable
Else
  Not acceptable
End If
Sustainability Phase 2010-2013

Code released in September 2010
Production distribution in December 2010
Project partners providing self-funded maintenance (but not development)

Three years of training & support functions planned by project partners

By 2014, transitioned long-term support of tool to an appropriate organization
Workshops & Trainings

Workshops proposed for 2011

- Code4Lib (Bloomington, Feb. 7-10)
- IS&T Archiving (Salt Lake City, May 16-19)
- Open Repositories (Austin, June 8-11)

Anticipate more trainings, more vehicles

- Webinars & videos
Future & 3rd Party Development

Known 3rd party development activities

– Integration with DuraCloud (DuraSpace)
– ARC module for Web Archiving (Bibliothèque nationale de France)
– WARC, JPEG, GIF modules (CDL, hopefully ;-)

Possible development efforts

– Additional format modules: QuickTime, MXF, AVI?
– Configuration GUIs
– Integration with DSpace, Fedora, FITS, etc.

Suggestions, volunteers and funders welcome!
Questions?

http://jhove2.org
JHOVE2-Announce-L@listserv.ucop.edu
JHOVE2-Techtalk-L@listserv.ucop.edu

CDL
Stephen Abrams
Patricia Cruse
John Kunze
Isaac Rabinovitch
Marisa Strong
Perry Willett

Stanford University
Richard Anderson
Tom Cramer
Hannah Frost

Portico
John Meyer
Sheila Morrissey

Library of Congress
Martha Anderson
Justin Littman

Advisory Board
Bibliothèque national de France
Deutsche Nationalbibliothek
DSpace / MIT
Ex Libris
Fedora Commons / Rutgers
Florida Center for Library Automation
Harvard University
Koninklijke Bibliotheek
National Archives (UK)
National Archives (US)
National Library of Australia
National Library of New Zealand
Planets / Universität zu Köln
Tessella