Next-Generation Characterization
An Update on the JHOVE2 Project

JHOVE2 Project Team
California Digital Library, Portico, Stanford University
The preservation problem

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

“Tell me about yourself...”
“What? So what?”

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

- Identification
- Feature extraction
- Validation
- Assessment

“What is it?”
“What about it?”
“What is it, really?”
“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
Characterization in migration workflows
JHOVE2 project

Build on the success of JHOVE, addressing some of its known deficiencies of design and implementation, and extending its function

- Collaboration of CDL, Portico, and Stanford
- Funded by NDIIPP
- Open source deliverables (BSD)
Feature set

Multi-stage processing

- Signature-based identification
  - DROID
- Feature extraction
- Validation
- Message digesting
  - Adler-32, CRC-32, MD2, MD5, SHA-1, SHA-256, SHA-384, SHA-512
- Rules-based assessment

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

Recursive processing of objects arbitrarily-nested within containers
Feature set

Granular modularization with generic plug-ins
Clean APIs and common module design patterns
Buffered I/O
Internationalized output
Extensive configuration via dependency injection
Complete documentation
   – User’s guide
   – Architectural overview
   – Module specifications
   – Programmer’s guide
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 (ICC.1:2004-10)
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, ...
UTF-8 ASCII (ANSI X3.4)
WAVE BWF (EBU N22-1997)
XML
Zip
Supported formats

**netCDF**
[http://www.unidata.ucar.edu/software/netcdf](http://www.unidata.ucar.edu/software/netcdf)

**Grib**

- Developed by the Wegener Institute (Germany)
  [http://www.awi-potsdam.de](http://www.awi-potsdam.de)
- Widely used for meteorological data
(Un)supported formats

AIFF
GIF
HTML
JPEG

– HTML can be expressed in terms of SGML or XML
– We’re investigating funding options for subsequent development of GIF and JPEG modules
Implementation

Java 1.6 J2SE
http://java.sun.com/javase/6/docs/api

- Annotations

- Buffered I/O (java.nio)
  http://java.sun.com/javase/6/docs/api/java/nio/package-summary.html

- Reflection
  http://java.sun.com/docs/books/tutorial/reflect

Spring dependency injection framework
http://www.springframework.org/

Mercurial distributed code repository
http://mercurial.selenic.com/

Maven build management
http://maven.apache.org/

Bitbucket code hosting
http://www.bitbucket.org/
Properties and reportables

A *property* is a named, typed value

- Name
- Unique formal identifier
- Data type
  - Scalar or collection
  - Java types, JHOVE2 primitive types, or JHOVE2 *reportables*
- Typed value
- Description of correct semantic interpretation

A *reportable* is a named set of properties

- Reportables correspond to Java *classes*
- Properties correspond to *fields*
Source units

A formatted object about which characterization information can be meaningfully reported

- **Unitary**
  - File
  - File inside of a container
  - Byte stream inside a file
  - e.g. TIFF
  - e.g. TIFF inside a Zip
  - e.g. ICC inside a TIFF

- **Aggregate**
  - Directory
  - Directory inside of a container
  - File set
  - Clump
  - e.g. command line arguments
  - e.g. Shapefile

For purposes of characterization, directories, file sets, and clumps are considered formats
Characterization strategy

1. Identify format
2. Dispatch to appropriate format module
   a) Extract format features and validate
      – If a nested source unit is found, process recursively, (go to Step 1)
   b) Validate format profiles (optional)
3. If unitary source unit, calculate message digests
4. Assess
5. If aggregate source unit, try to identify aggregate format, and if successful, process recursively (go to Step 1)
Characterization strategy
Characterization strategy

- abc.shp
- abc.shx
- abc.dbf
- abc.tif
- xyz.pdf

Main
Index
dBASE
GeoTIFF
PDF
Characterization strategy
Characterization strategy

- abc.shp
- abc.shx
- abc.dbf
- abc.tif
- "GIS object"
- clump
- GeoTIFF
- Shapefile
- Main
- Index
- dBASE
- xyz.pdf
- PDF
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

Assessment rules are logical expressions of the form

If *condition* then *consequent* else *alternative*

– A condition is defined by either a universal or existential qualifier
  \[ \forall \text{ “for all”} \]
  \[ \exists \text{ “there exists” or “for any”} \]

and an arbitrary set of predicates (logical assertions) of the form

*property relation value*

– Supported relational operators
  \[ = \quad \neq \quad < \quad > \quad \leq \quad \geq \quad \text{contains} \]
Assessment

XML rule example (pseudocode)

If ALL_OF
    xmlDeclaration.standalone == 'yes'
    valid.toString() == 'true'
Then
    Acceptable
Else
    Not acceptable
End If

Predicates are evaluated using MVEL

http://mvel.codehaus.org/
Demonstration

% jhove2 [-ik] [-b size]
  [-B Direct|NonDirect|MAPPED]
  [-d JSON|Text|XML] [-f limit]
  [-t temp] [-o file] file ...

- **i**  Show identifiers in JSON and Text displayers
- **k**  Calculate message digests
- **-b size**  I/O buffer size, in bytes  \( \text{(default: 131072)} \)
- **-B type**  I/O buffer type: Direct, NonDirect, Mapped  \( \text{(default: Direct)} \)
- **-d displayer**  Displayer: JSON, Text, XML  \( \text{(default: Text)} \)
- **-f limit**  Fail fast limit  \( \text{(default: 0, no limit)} \)
- **-t temp**  Temporary directory
- **-o file**  Output file  \( \text{(default: standard output)} \)
- **file**  File or directory
User survey

145 respondents, 88 institutions, 23 countries

1) Please characterize your institution:

- National library / Archives: 22%
- Academic library / archives: 49%
- Government agency: 7%
- Non-profit org: 9%
- For-profit org: 6%
- Library consortium: 2%
- Public library / archives: 5%

3) How quickly do you plan to begin using JHOVE2 after its release?

- 0-6 months: 64%
- 7-12 months: 18%
- 13-18 months: 2%
- Not sure: 18%
- Other: 2%

Full results available at https://confluence.ucop.edu/display/JHOVE2Info/User+survey
User survey

5) Please characterize how you will use JHOVE2

- Integral part of validation process for all digital objects we ingest into our repository (36%)
- Integral part of process for all digital objects we create (20%)
- Integral part of quality assurance process for all digital objects we receive from outside (24%)
- Ad hoc use (14%)
- Not sure (8%)
- Other (3%)

6) What resources would be most helpful in adopting JHOVE2

- Written documentation (38%)
- Reference implementation (28%)
- Webinar (15%)
- Face-to-face training session / tutorial (6%)
- Conference presentation (6%)
- Implementation fail other (2%)

Full results available at https://confluence.ucop.edu/display/JHOVE2Info/User+survey
Sustainability

Final production release in September 2010

Workshop at iPRES 2010, Vienna, September 19-24
http://www.ifs.tuwien.ac.at/dp/ipres2010

Project partners will provide ongoing, self-funded maintenance (but not development)

Funded development activities

- Integration with DuraCloud (DuraSpace)
- ARC and WARC modules (Bibliothèque nationale de France)
Sustainability

Possible development efforts

- Additional format modules
- Configuration GUIs
- JHOVE2-as-a-service
- Integration with
  ✓ DAITTS, DSpace, Fedora, FITS, etc.

Training and tutorials

- “Train the trainer”

Look for a permanent institutional home
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

Portico
John Meyer
Sheila Morrissey

Library of Congress
Martha Anderson
Justin Littman

Stanford University
Richard Anderson
Tom Cramer
Hannah Frost

With help from
Walter Henry
Nancy Hoebelheinrich
Keith Johnson
Evan Owens

Advisory Board
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