UN Food and Agriculture Organization, Rome, 23-26 May 2011

JHOVE2 Next-Generation Characterization Workshop

Stephen Abrams
Perry Willett
California Digital Library

Sheila Morrissey
Portico

Tom Cramer
Stanford University
Agenda

UN Food and Agriculture Organization
Viale delle Terme di Caracalla, 00153 Rome, Italy
23-26 May 2011

Austria Room
• Day 1 – Introduction to digital preservation
• Day 2 – Preservation case studies and introduction to characterization

Ethiopia Room
• Day 3 – JHOVE2 concepts, installation, and configuration
• Day 4 – Community building and sustainability, and JHOVE2 module development
## Day 2 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:30 am</td>
<td>Infrastructure and tools</td>
</tr>
<tr>
<td>09:30 – 10:30 am</td>
<td>Case study: preservation activities at CDL</td>
</tr>
<tr>
<td>10:30 – 11:00 am</td>
<td>Break</td>
</tr>
<tr>
<td>11:00 – 12:00 pm</td>
<td>Case study: preservation activities at Portico</td>
</tr>
<tr>
<td>12:00 – 12:30 pm</td>
<td>Preservation initiatives and organizations</td>
</tr>
<tr>
<td>12:30 – 14:00 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Case study: preservation activities at Stanford</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:10 pm</td>
<td>Automated characterization</td>
</tr>
<tr>
<td>16:10 – 16:30 pm</td>
<td>Format characterization in preservation workflows</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
Characterization

- Preservation management is concerned with the gap between what you were given (in the past) and what you need (in the future)
  - 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


How does what I have (now) compare to what I had (before)?

How does what I have (now) compare to what I want (ahead)?
“Tell me about yourself...”

Manual characterization is not feasible at any significant scale.

Automation facilitates characterization at scale by programmatically examining digital objects for properties that can be extracted or inferred.

© United Features Syndicate, Inc.
Characterization

• How do you know what you have?

• How can you confirm you received what you expected?
  – Less than a third of respondents to the 2009 Planets survey felt they had “control” over decisions regarding what content they were being asked to manage

  *Survey Analysis Report, IST-2006-033789, DT11-D1, 2009-05-06*

• How will you classify for purposes of analysis, planning, and efficient workflow?
  – Categorization to facilitate highly automated workflows
  – Treat like objects alike
Characterization

• OAIS representation information
  – What you need to know in order to interpret a content object properly (ISO 14721:2003)

• Significant properties
  – “Those characteristics (both technical, intellectual, and aesthetic) agreed by the archive or by the collection manager to be the most important features to preserve over time” (Cedars project, 2001)
Characterization

• Descriptive
  – What is the *intellectual* description of the content? What is the content *about*?

• Administrative
  – What are the properties necessary to manage this content object? Who is its owner? Who is its curator? Who pays the bills?

• Structural
  – What are the *relationships* between the various components that make up a content object?

• Format
  – What are the *technical* properties defined by the object’s format?
Characterization

• Why worry about formats?

*Preservation of information*

Format

*Preservation of bits*

• Since formatted digital assets are inherently mediated by technology, they are particularly susceptible to disruptive technological change

A set of syntactic and semantic rules for mapping between an information model and a serialized bit stream
Characterization

- Without strong format typing, all content is opaque
Format characterization

• The automated determination of the intrinsic and extrinsic properties of a formatted object
  
  Identification
  Feature extraction
  Validation
  Assessment

Determining the presumptive format of a digital object based on suggestive extrinsic hints and intrinsic signatures

C:\My Documents\first-day-of-creation.jpg

```
ffd8ffe000104a46494600010201008300830000ffed0fb050686f ...”
```

“This JPEG is at 600 dpi”

“This JPEG does not have a required Start-of-Image segment”

Determining the level of conformance to the normative requirements of a format’s authoritative specification

Determining the level of acceptability for a specific purpose on the basis of locally-defined policy rules

“We report, you decide”
Validation vs. assessment

• A perfectly valid object may not be acceptable
  – Reformatting outputs may not conform to 600 dpi, sRGB, lossless compression, etc.

• An invalid object may be (grudgingly) acceptable
  – Many TIFF images are technically invalid but are renderable
  – Some PDF documents are technically invalid (even those produced by Adobe tools!) but are renderable
  – Most HTML pages are technically invalid but are renderable
  – Different tools may recover from validation errors in different ways; permissive tools encourage bad practice
Format profiles

• Many formats define a “family” of digital encoding schemes
  – TIFF (Tagged Image File Format)
    • Big-endian vs. little-endian
    • Version 4, 5, 6
    • Class B, F, G, P, R, Y
    • TIFF/EP (ISO 12234-2)
    • TIFF/IT (ISO 12639)
      – BP, BL, CT, FP, HC, SD
      – P1, P2
    • GeoTIFF
    • DNG (Digital Negative)

• Being able to distinguish between these profiles may be significant for purposes of analysis and planning
Characterization tools

• Genre-specific tools
  – E.g. ImageMagick for images
  – Exiftool for Exif images (specific versions of TIFF and JPEG)
  – PDF “pre-flight”

• Forensic tools

• Unix “file” utility

• National Archives (UK) DROID

• National Library of New Zealand metadata extractor

• JHOVE(1)

• Harvard University FITS (File information tool set)

• JHOVE2
Demonstration
Characterization in ingest workflows
Characterization in migration workflows

- AIP
- Unpackage
- Assessment
- Migration
- Equivalence
- (Re)Ingest

- Content
- Content'
- Identification
- Feature extract
- Validation
- Equivalence

- Policy rules
- Metadata
- Metadata'

- Content

UN FAO Workshop
Characterization summary

• Characterization is the automated process of extracting or inferring the properties of a formatted digital object significant for purposes of classification, analysis, and planning

• An understanding of format is important to facilitate preservation of information, as opposed to preservation of bits

• Introduce characterization as far upstream in the ingest process as possible

• Always perform before/after characterization whenever introducing changes to content state
Questions? Discussion?

For more information...

http://jhove2.org/

JHOVE2-Announce-L@listserv.ucop.edu
JHOVE2-Techtalk-L@listserv.ucop.edu

CDL/UC3
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

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

UN FAO Workshop
# Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
Objectives

• Understand the role of characterization, including identification, feature extraction, validation, and assessment, in digital curation and preservation workflows

• Appreciate the functionality of the JHOVE2 application, including the significant enhancements relative to JHOVE, and new capabilities based on object- and aggregate-level characterization

• Learn the architecture, components, design patterns and API’s of the JHOVE2 framework, as well as the configuration options for plug-in modules, characterization strategies, and results formatting

• Demonstrate the use of JHOVE2’s new rule-based assessment capabilities, and integrating these into local workflows to determine object acceptability

• Gain a better understanding of the community model for the project and how individual institutions can contribute new format modules as well as resources to help extend and sustain the open source project
But first, a few questions...

• Are there people here today that did *not* attend the introduction to characterization on Tuesday afternoon?

• Who are you, where are you from, what are your local preservation activities, how do you expect to use JHOVE2?
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td><em>Review of objectives and agenda</em></td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td><strong>JHOVE2 project</strong></td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td><em>Break</em></td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td><em>Lunch</em></td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td><em>Break</em></td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
JHOVE2 project

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

• Collaboration between the California Digital Library (CDL), Portico, and Stanford University

• Funded by the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP)
Project goals

• Address known deficiencies in design and implementation of JHOVE(1)
  – API complexity and idiosyncrasy
    • Granular modularization with generic plug-ins
    • Standardized module design patterns
  – Internationalization
    • Java localization
  – Performance
    • Java buffered I/O
Project goals

• Provide enhancements to JHOVE2 functionality
  – Multi-stage processing
    • Signature-based identification
      – DROID for files
      – Pathname globbing for aggregates
    • Feature extraction
    • Validation
    • Message digesting
    • Rules-based assessment
  – Recursive processing of arbitrarily-nested objects
  – Support of complex objects spanning multiple files
Project goals

• Provide enhancements to JHOVE2 functionality
  – Extensive configuration via Spring dependency injection and Java properties files
    • Characterization strategy
    • Module customization
    • Message localization

• Complete documentation
  – User’s guide
  – Module specifications
  – Architectural overview
  – Programmer’s guide
Project goals

• Facilitate and encourage third-party maintenance and enhancement of the codebase
  – API simplification
  – Adherence to common module design patterns
  – Extensive documentation

• It’s working!
  – NetCDF and Grib modules (Wegener Institute)
  – Gzip and ARC modules (Bibliothèque nationale de France)
Supported formats

• JHOVE2 can identify (via DROID) many more formats than it can validate (via modules)
  – PRONOM registry documents over 550 formats
    [link to PRONOM]

[Table of file formats]
Supported formats

- Directory
- File set
- ICC color profile
- JPEG 2000  JP2, JPX
- PDF  1.0 – 1.7, ISO 3200-1, PDF/A, PDF/X
- SGML
- Shapefile  Main, Index, dBASE, ...
- UTF-8  ASCII
- WAVE  BWF
- XML
- Zip
Supported formats

Wegener Institute (Germany)
http://www.awi-potsdam.de/
• NetCDF
  http://www.unidata.ucar.edu/software/netcdf
• Grib

Bibliothèque nationale de France / Atos Origin
• Gzip
  http://www.gzip.org/zlib/rfc-gzip.html
• ARC
(Un)Supported formats

- **AIFF**
  All of these formats remain supported in JHOVE1

- **GIF**
  We’re investigating funding options for follow-on work to develop GIF and JPEG modules

- **HTML**
  HTML can be expressed in terms of SGML or XML

- **JPEG**
Implementation

• Java 1.6 J2SE
  http://java.sun.com/javase/6/docs/api
    – Annotations
    – Buffered I/O
      http://java.sun.com/javase/6/docs/api/java/nio/package-summary.html
    – Reflection
      http://java.sun.com/docs/books/tutorial/reflect

• Spring framework
  http://www.springframework.org/
    – Dependency injection (DI) / inversion of control (IOC)

• BerkeleyDB JE (Java edition)
Implementation

• Bitbucket code hosting
  http://www.bitbucket.org/
  – Maven
    http://maven.apache.org/
  – Mercurial
    http://mercurial.selenic.com/
JHOVE2 project

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

Portico
- John Meyer
- Sheila Morrissey

Stanford
- Richard Anderson
- Tom Cramer
- Hannah Frost

Library of Congress
- Martha Anderson
- Justin Littman

With help from
- Walter Henry
- Nancy Hoebelheinrich
- Keith Johnson
- Evan Owens

Advisory board
- Bibliothèque nationale de France
- Deutsche Nationalbibliothek
- Ex Libris
- Fedora Commons / Rutgers University
- Florida Center for Library Automation
- Harvard University / GDFR project
- Koninklijke Bibliotheek
- Library of Congress
- MIT / DSpace
- NARA
- National Library of Australia
- National Library of New Zealand
- Planets project / Universität Köln
- Tessella

With help from
- Walter Henry
- Nancy
- Hoebelheinrich
- Keith Johnson
- Evan Owens
JHOVE2 project

http://jhove2.org/

JHOVE2-Announce-L@listserv.ucop.edu
JHOVE2-Techtalk-L@listserv.ucop.edu
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
Reportable property

• A significant characteristic of a formatted digital object that can be extracted or inferred
  
  – Name          Property name as defined by the format
  – Type          Scalar or collection; Java and JHOVE2 types
  – Value
  – Unit          Optional unit of measure label
  – Identifier    Unique JHOVE2 identifier
  – Description   Optional description of property semantics
  – Reference     Optional reference to the controller section of the format specification

• A “Reportable” is a named aggregation of reportable properties
  
  – A Reportable is represented by a Java class; a reportable property, by a field and methods in that class
Reportable property

• Almost all conceptual entities are represented by Reportables

• The output from JHOVE2 is structured as a hierarchy of Reportables
  
  – This hierarchical structure is implied by indentation in the Text handler; it is explicit in the nesting structure of the JSON and XML handlers
Identifier

• All JHOVE2 entities are associated with unique identifiers

    http://jhove2.org/terms/type/specific

where *type* indicates the general type

    format, message, property, reportable

and *specific* indicates the specific entity

• Format identifiers are based on the format common name

    http://jhove2.org/terms/format/utf-8
Identifier

• Reportable identifiers are based on the underlying class name
  
  http://jhove2.org/terms/reportable/org/jhove2/core/JHOVE2

• Property and message identifiers are based on the underlying class name and accessor method
  
  http://jhove2.org/terms/property/org/jhove2/core/JHOVE2/Commands
  http://jhove2.org/terms/message/org/jhove2/module/utf8/UTF8Module/ByteOrderMark

• Implemented as a Reportable
Source unit

• Any digital entity that can be meaningfully characterized
  – A file or web resource (i.e. something with a single URL)
  – A byte stream within a file (or web resource)
  – A collection of files (i.e. a PREMIS Representation)

• Implemented as a Reportable

• May encapsulate subsidiary source units
  – All children of a given parent are automatically recursively characterized during the processing of the parent
  – If the children may be arbitrary and have no obvious intellectual relationship to the parent, the parent is considered aggregate; otherwise it is unitary
Source unit

DirectorySource

PresumptiveFormats
DirectoryModule

ChildSources

FileSource

PresumptiveFormats
FormatModule

ChildSources

ByteStreamSource

ByteStreamSource

FileSource

PresumptiveFormats
FormatModule
Source unit

• Explicitly aggregate source units
  – Directory *File system or container file directory*
  – File set *The set of objects specified on the command line*

• Explicitly unitary source unit
  – Byte stream

• Initially assumed unitary, but may be determined aggregate during processing
  – File
  – URL
Source unit

• Aggregate source units are subject to an extra processing step known as *aggrefication* (i.e. *aggregate identification*)

• If the aggregate holds a coherent characterizable entity, a new source unit known as a *Clump* is inserted into the source hierarchy
Source unit

• Common source unit properties
  – Backing file
  – Children
  – File system properties
  – Extra properties
  – isAggregate
  – Messages
  – Modules
  – Presumptive format(s)
Message

• A source unit may be associated with messages documenting various conditions
  – Error \quad A terminal condition requiring remedial action
  – Warning \quad A condition possibly requiring attention
  – Informative \quad A condition not requiring further attention

• Messages can arise in two contexts
  – Process \quad An unanticipated condition arising from the characterization process
  – Object \quad An unexpected condition in the characterized source

• Implemented as a Reportable
Characterization strategy

• The iterative sequence of processing steps applied to every source unit
  – Identify format *(if not previously identified)*
  – Dispatch to format module
    • Extract features and validate
      – If nested source unit discovered, *process* recursively...
    • Validate format profiles *(if registered)*
  – If unitary, *calculate* message digests *(optional)*
  – Assess
  – If aggregate, *aggregate* identification
    • If a Clump, *process* recursively...
Input

• An abstraction used to support uniform access to source units regardless of their underlying data structure

• Based on java.nio buffers
  – Direct \textit{I/O subsystem of underlying OS}
  – Non-direct \textit{I/O subsystem of JVM}
  – Memory mapped \textit{Paging subsystem of underlying OS (1.6 GB limit)}

\begin{center}
\begin{tikzpicture}
\draw[->] (-3,0) -- (3,0);
\draw[->] (0,-1) -- (0,1);
\node at (-3,-0.5) {Non-direct};
\node at (0,-0.5) {Direct};
\node at (3,-0.5) {Memory mapped};
\node at (-3,1) {Fastest initialization}\node at (0,1) {Slowest performance}\node at (3,1) {Fastest performance};
\node at (-3,-1) {Slowest performance}\node at (0,-1) {Slowest initialization}\node at (3,-1) {Fastest performance};
\end{tikzpicture}
\end{center}
Persistence

• JHOVE2 builds an in-memory representation of characterization information (a hierarchy of Reportables)
  – If invoked against a large set of source units, the memory footprint will grow correspondingly large, to the point of a JVM out-of-memory error

• JHOVE2 supports characterization of arbitrarily-large source unit sets in a fixed memory footprint
  – BerkeleyDB JE (Java Edition), an open source, embeddable No-SQL database
  – Use of BDB JE is a configurable option
Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td><strong>Review of objectives and agenda</strong></td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td><strong>JHOVE2 project</strong></td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td><strong>Concepts</strong></td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td><strong>Demonstration</strong></td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td><strong>Architecture</strong></td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td><strong>Lunch</strong></td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td><strong>Installation</strong></td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td><strong>Configuration</strong></td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td><strong>Community building and sustainability</strong></td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td><strong>Questions and discussion</strong></td>
</tr>
</tbody>
</table>
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td><strong>Review of objectives and agenda</strong></td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td><strong>JHOVE2 project</strong></td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td><strong>Concepts</strong></td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td><strong>Demonstration</strong></td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td><strong>Lunch</strong></td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
# Command line invocation

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

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-i</td>
<td>--show-identifiers</td>
</tr>
<tr>
<td>-k</td>
<td>--calc-digests</td>
</tr>
<tr>
<td>-b size</td>
<td>--buffer-size size</td>
</tr>
<tr>
<td>-B type</td>
<td>--buffer-type type</td>
</tr>
<tr>
<td>-d display</td>
<td>--display display</td>
</tr>
<tr>
<td>-t temp</td>
<td>--temp temp</td>
</tr>
<tr>
<td>-o out</td>
<td>--output out</td>
</tr>
<tr>
<td>file</td>
<td></td>
</tr>
</tbody>
</table>
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
Architecture

• Three conceptual layers
  – Application  Command-line application
  – Framework  Coordinates all processing
  – Modules    Embodies specific processing behaviors

• Technically, the JHOVE2 application and framework are implemented as modules, however, due to their central, but distinct, roles in processing, it is useful to consider them as conceptually independent levels
Framework

- Coordinates all JHOVE2 processing
- Invokes the configured *characterization strategy* for all source units
Modules

• All JHOVE2 behaviors are embodied by modules
  – Application
  – Framework
  – Commands
  – Strategy modules
  – Format modules
  – Format profiles
  – Displayers
Commands and strategy modules

• The framework is configured to invoke command modules, which in turn invoke strategy modules
  – IdentifierCommand $\Rightarrow$ IdentifierModule
    • Signature-based identification using DROID
  – DispatcherCommand $\Rightarrow$ format-specific module
  – DigesterCommand $\Rightarrow$ DigesterModule
    • Adler-32, CRC-32, MD2, MD5, SHA-1, SHA-256, SHA-384, SHA-512
  – AssessmentCommand $\Rightarrow$ AssessmentModule
  – AggrefierCommand $\Rightarrow$ AggrefierModule
    • Pathname globbing identification
Format modules

• A format module must be capable of parsing a formatted source unit and extracting its pertinent features

• It may be capable of validating the conformance of the source unit to the normative rules of its format

• A format module will attempt to fully parse the source unit, even after it is determined to be invalid
Format profiles

• A format profile is a subtype of a format
  – ASCII is a profile of UTF-8
  – BWF is a profile of WAVE
  – TIFF/EP is a profile of TIFF

• Profile validation is automatically performed
  – Profiles do not reparse the source unit; validation is based on previously extracted features

• Profile validity is always reported, whether true or false
  – If false, the profile will report the features that are out of conformance
Architectural overview

https://bitbucket.org/jhove2/main/wiki/Architecture
# Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
# Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td><strong>Review of objectives and agenda</strong></td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td><strong>JHOVE2 project</strong></td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td><strong>Concepts</strong></td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td><strong>Demonstration</strong></td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td><strong>Architecture</strong></td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td><strong>Lunch</strong></td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td><strong>Installation</strong></td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td><strong>Configuration</strong></td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td><strong>Community building and sustainability</strong></td>
</tr>
<tr>
<td>16:30 – 17:00 pm</td>
<td><strong>Questions and discussion</strong></td>
</tr>
</tbody>
</table>
Installation

• Prerequisites
  – Java 1.6 JRE or JDK
  – Web browser (to download distribution package)
  – Zip or gunzip/tar utilities (to disaggregate package)
  – 68 MB file system
Download

https://bitbucket.org/jhove2/main/downloads
Disaggregate the Zip or tar.gz
Installation directory structure
Installation directory structure
Installation directory structure
Installation directory structure
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td><strong>Configuration</strong></td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:45 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
Invocation scripts

• All scripts are paired, one each for Unix (.sh) and Windows (.cmd)

  – arules.cmd
    arules.sh  Assessment rule generator
  – env.cmd
    env.sh  Java environment definition
  – jhove2.cmd
    jhove2.sh  JHOVE2 command line application
  – jhove2_doc.cmd
    jhove2_doc.sh  Module documentation generator
  – jhove2_dpfg.cmd
    jhove2_dpfg.sh  Displayer configuration generator
  – jhove2_upfg.cmd
    jhove2_upfg.sh  Units of measure configuration generator
Invocation environment

• The file env.sh (or env.cmd) defines the invocation environment

• In most cases the default environment will work without modification

  – JAVA_HOME              Undefined, script searches execution path
  – JAVA                  java or JAVA_HOME/bin/java
  – JHOVE2_HOME           Installation directory
  – CP (classpath)        All jar files in JHOVE2_HOME/lib
Configuration

- DROID configuration
- Message localization
- Java properties files
- Spring configuration
Message localization

# messages.properties
# Key value pairs from fully-qualified Java path for a Message field
# in a class to message text template Used for localization
#
# #############################################################################
# Message templates for class org.jhove2.core.JHOVE2
# #############################################################################
#
org.jhove2.core.JHOVE2.FileNotFoundMessage=File or directory not found:\ {0}
org.jhove2.core.JHOVE2.FileNotReadableMessage=File or directory not readable
#
# #############################################################################
# Message templates for class org.jhove2.module.aggregfy.AggregfierCommand
# #############################################################################
#
org.jhove2.module.aggregfy.AggregfierCommand.IOException=IOException thrown
...

config/messages/jhove2_messages.properties
Displayer directives

# _displayer.properties
# The visibility directives control the display of the properties identified
# by URI. The directives can be: Always, IfFalse, IfNegative, IfNonNegative,
# IfNonPositive, IfNonZero, IfPositive, IfTrue, IfZero, Never
# A property is not displayed if its value is not consistent with the
# directive.
# Negative means ..., -2, -1; NonNegative means 0, 1, 2...
# Positive means 1, 2, 3,...; NonPositive means ..., -2, -1, 0
http://jhove2.org/terms/property/org/jhove2/core/JHOVE2/Commands Always
http://jhove2.org/terms/property/org/jhove2/core/JHOVE2/Installation Always
http://jhove2.org/terms/property/org/jhove2/core/JHOVE2/Invocation Always
http://jhove2.org/terms/property/org/jhove2/core/JHOVE2/MemoryUsage Always
http://jhove2.org/terms/property/org/jhove2/core/JHOVE2/SourceCounter Always

config/properties/module/display/displayer/org/jhove2/...

reportable_displayer.properties

cfg/properties/module/display/displayer/org/jhove2/core/

JHOVE2_displayer.properties
Units of measure labels

# Units of measure properties
# Note: These unit of measure labels are descriptive only; changing the
# label does NOT change the determination of the underlying property value.
http\://jhove2.org/terms/property/org/jhove2/core/JHOVE2/MemoryUsage byte

config/properties/module/display/units/org/jhove2/...
    reportable_unit.properties

config/properties/module/display/units/org/jhove2/core/
    JHOVE2_unit.properties
Spring configuration

• Characterization strategy
  config/spring/jhove2-framework-config.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<beans xmlns="http://www.springframework.org/schema/beans" ... >
  ...
  <bean id="JHOVE2" class="org.jhove2.core.JHOVE2" scope="prototype">
    <constructor-arg ref="FrameworkAccessor"/>
    ...
    <property name="commands">
      <list value-type="org.jhove2.module.Command">
        <ref bean="IdentifierCommand"/>
        <ref bean="DispatcherCommand"/>
        <ref bean="DigesterCommand"/>
        <ref bean="AssessorCommand"/>
        <ref bean="AggregferCommand"/>
      </list>
    </property>
    ...
  </bean>
  ...
</beans>
```
Spring configuration

- Digester algorithms

```xml
<bean id="DigesterModule" class="org.jhove2.module.digest.DigesterModule" ... >
  ...
  <property name="arrayDigesters">
    <list value-type="org.jhove2.module.digest.ArrayDigester">
      <!-- <ref bean="Adler32Digester"/> -->
      <ref bean="CRC32Digester"/>
    </list>
  </property>
  <property name="bufferDigesters">
    <list value-type="org.jhove2.module.digest.BufferDigester">
      <!-- <ref bean="MD2Digester"/> -->
      <ref bean="MD5Digester"/>
      <ref bean="SHA1Digester"/>
      <!-- <ref bean="SHA256Digester"/> -->
      <!-- <ref bean="SHA384Digester"/> -->
      <!-- <ref bean="SHA512Digester"/> -->
    </list>
  </property>
</bean>

<bean id="Adler32Digester" class="org.jhove2.module.digest.Adler32Digester" ... >
  ...
```
Spring configuration

- Persistence manager
  
  config/spring/persist/jhove2-persist-config.xml

```xml
<!-- Beans for in-memory persistence -->
<!--
<bean id="SourceFactory" class="org.jhove2.persist.inmemory.InMemory..." />
<bean id="ApplicationModuleAccessor" ... />
...
<bean id="BaseModuleAccessor" ... />
 -->

<!-- Beans for BerkeleyDB persistence -->
<!-- -->
<bean id="SourceFactory" class="org.jhove2.persist.berkeleydpl.BerkeleyDb..." />
<bean id="ApplicationModuleAccessor" ... />
...
<bean id="BaseModuleAccessor" ... />
<!-- -->

<bean id="BerkeleyDbPersistenceManager" class="org.jhove2.persist.berkeleydpl...">
  <property name="envHome" value="C:\"/>
...
</bean>
```
Spring configuration

• Persistence manager
  properties/persistence/persistence.properties

```java
# classname org.jhove2.config.spring.SpringInMemoryPersistenceManagerFactory
classname org.jhove2.config.spring.SpringBerkeleyDbPersistenceManagerFactory
```
User’s guide
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:45 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:45 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Community building and sustainability</td>
</tr>
<tr>
<td>16:45 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
## Day 3 agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 – 08:35 am</td>
<td>Review of objectives and agenda</td>
</tr>
<tr>
<td>08:35 – 09:15 am</td>
<td>JHOVE2 project</td>
</tr>
<tr>
<td>09:15 – 10:15 am</td>
<td>Concepts</td>
</tr>
<tr>
<td>10:15 – 10:45 am</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 – 11:30 am</td>
<td>Demonstration</td>
</tr>
<tr>
<td>11:30 – 12:00 pm</td>
<td>Architecture</td>
</tr>
<tr>
<td>12:00 – 13:30 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30 – 14:00 pm</td>
<td>Installation</td>
</tr>
<tr>
<td>14:00 – 15:00 pm</td>
<td>Configuration</td>
</tr>
<tr>
<td>15:00 – 15:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>15:30 – 16:30 pm</td>
<td>Assessment</td>
</tr>
<tr>
<td>16:45 – 17:00 pm</td>
<td>Questions and discussion</td>
</tr>
</tbody>
</table>
Post-project planning

• Production 2.0.0 release in April 2011
• Additional releases scheduled for later in 2011
  – 2.1.0 ARC, Gzip, JPEG 2000, DROID 6
  – 2.2.0 PDF, Zip (fully validating)
• Project partners will provide ongoing, self-funded, post-release support and maintenance (but not development) for three years
• By year four, project partners expect to have transitioned long-term support, maintenance, and develop coordination to a permanent organization
Sustainable activities

• Support and maintain the core JHOVE2 code
• Provide training on integration and use
• Solicit and support 3rd party module development
• Solicit and support integration with other systems
• Establish a lightweight community structure to guide and foster JHOVE2 technical development
• Suggestions welcome, volunteers encouraged
Community

• Steering group: the three project partners, Library of Congress, other contributors as appropriate (based on level of commitment), all dedicated to sustaining the project and codebase

• Advisory group: providing strategic input and resources for maintenance and enhancement (based on vested interest)

• Committers group: JHOVE2 core developers (based on experience), advancing the core codebase, integrating contributions, and managing releases

• Open source community: JHOVE2 users and code contributors
User-driven priorities

- Planned activities are based on user survey results
  - 145 respondents, 88 institutions, 23 countries

User-driven priorities

• Planned activities are based on user survey results
  – 145 respondents, 88 institutions, 23 countries

Future and third-party development

• 3rd party development activities
  – NetCDF and Grib modules (Wegener Institute)
  – ARC and Gzip module (Bibliothèque nationale de France / Atos)
  – Integration with DuraCloud (DuraSpace)
  – WARC and HTML modules, virus detection
  – AIFF, JPEG, and GIF modules

• Possible development efforts
  – Additional format modules
  – Configuration GUIs
  – JHOVE2-as-a-service
  – Integration with DAITTS, DSpace, Fedora, FITS, etc.

• Suggestions, volunteers, and funders welcome!
Group discussion

• What training needs and opportunities do you see?

• Are there any particular modules that you think are critical priorities?

• Are there any that you’d like to develop?

• Are there any particular integrations you feel would be helpful in driving the adoption, utility, or enhancement of JHOVE2?

• Can you suggest any projects, funders, or opportunities that should be considered?
Questions? Discussion?

For more information...

http://jhove2.org/

JHOVE2-Announce-L@listserv.ucop.edu
JHOVE2-Techtalk-L@listserv.ucop.edu

CDL/UC3
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

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