JHOVE2

UC Curation Center CDL California Digital Library

PORTICO

Stanford University LIBRARIES & ACADEMIC INFORMATION RESOURCES

NDIIPP Partners Meeting

Arlington, Virginia, July 20-22, 2010

Next-Generation Characterization
An Update on the JHOVE2 Project

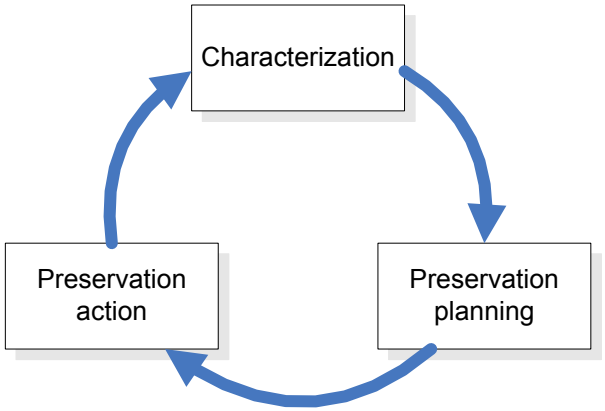
JHOVE2 Project Team

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



Adopted from A. Brown, "Developing Practical Approaches to Active Preservation," *IJDC* 2:1 (June 2007): 3-11.

"Tell me about yourself…"

"What? So what?"

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

- Identification — "What is it?"
- Feature extraction — "What about it?"
- Validation — "What is it, really?"
- Assessment — "So what?"

Validation vs. assessment

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

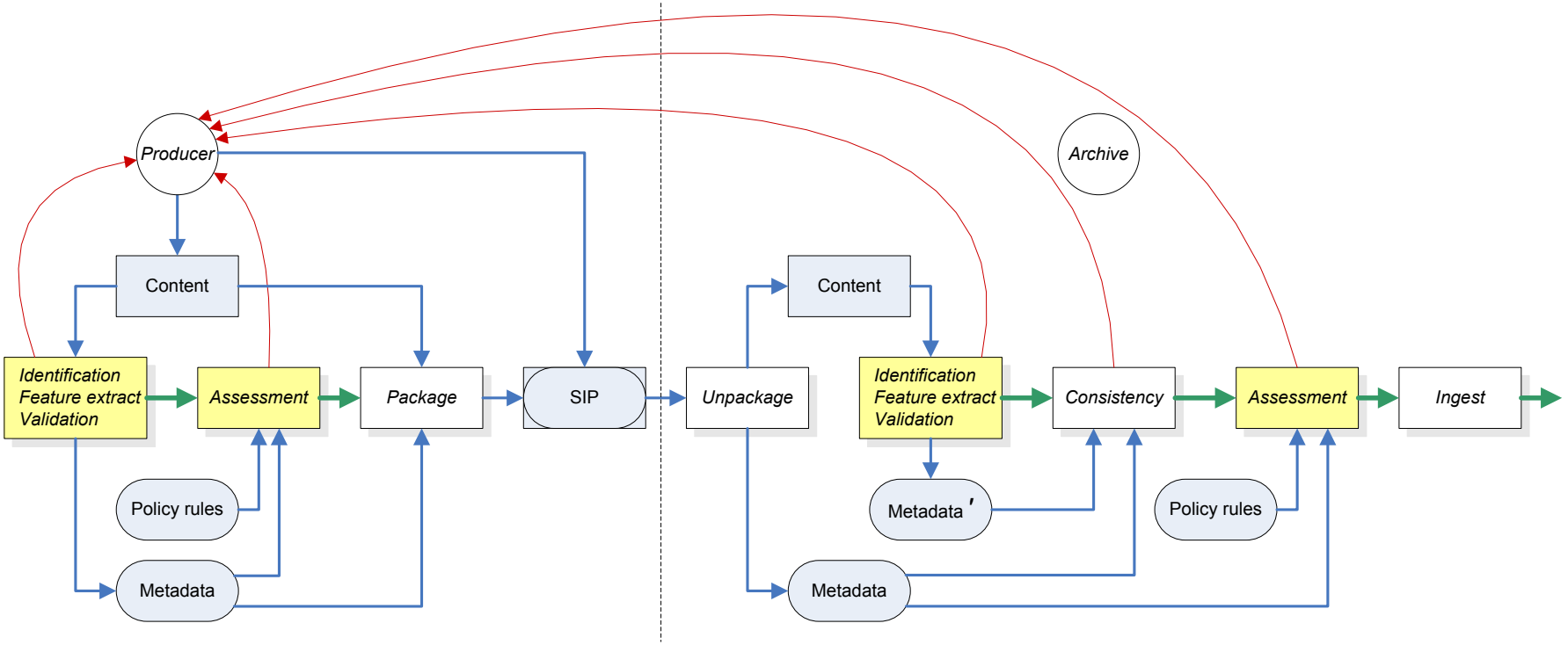
- To the extent that there is community consensus on these requirements, validation is an *objective* determination

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

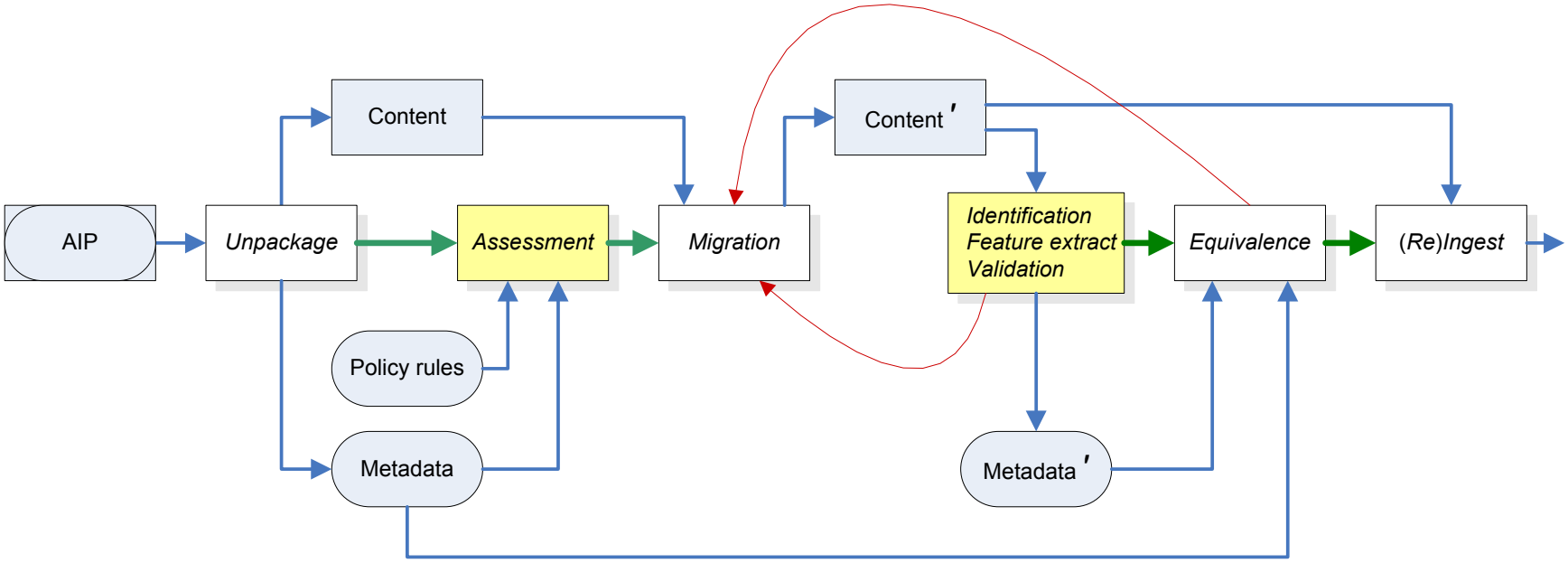
- Since these rules are locally configurable, assessment is a *subjective* determination

"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
    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

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

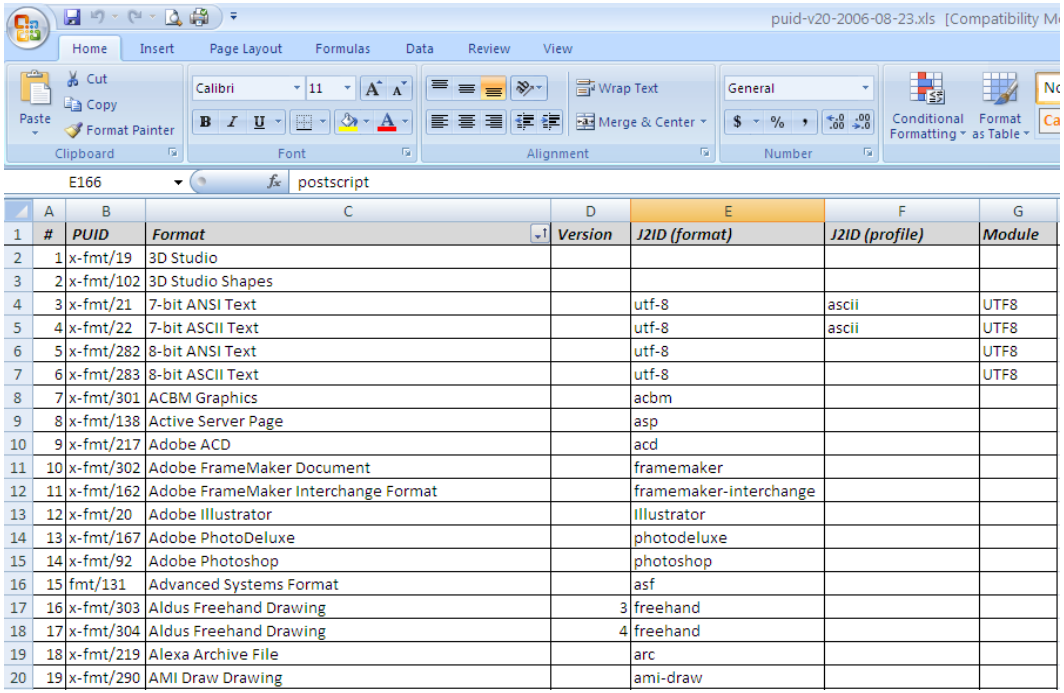
Complete documentation

- 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



	A	B	C	D	E	F	G
1	#	PUID	Format	Version	J2ID (format)	J2ID (profile)	Module
2	1	x-fmt/19	3D Studio				
3	2	x-fmt/102	3D Studio Shapes				
4	3	x-fmt/21	7-bit ANSI Text		utf-8	ascii	UTF8
5	4	x-fmt/22	7-bit ASCII Text		utf-8	ascii	UTF8
6	5	x-fmt/282	8-bit ANSI Text		utf-8		UTF8
7	6	x-fmt/283	8-bit ASCII Text		utf-8		UTF8
8	7	x-fmt/301	ACBM Graphics		acbm		
9	8	x-fmt/138	Active Server Page		asp		
10	9	x-fmt/217	Adobe ACD		acd		
11	10	x-fmt/302	Adobe FrameMaker Document		framemaker		
12	11	x-fmt/162	Adobe FrameMaker Interchange Format		framemaker-interchange		
13	12	x-fmt/20	Adobe Illustrator		Illustrator		
14	13	x-fmt/167	Adobe PhotoDeluxe		photodeluxe		
15	14	x-fmt/92	Adobe Photoshop		photoshop		
16	15	fmt/131	Advanced Systems Format		asf		
17	16	x-fmt/303	Aldus Freehand Drawing	3	freehand		
18	17	x-fmt/304	Aldus Freehand Drawing	4	freehand		
19	18	x-fmt/219	Alexa Archive File		arc		
20	19	x-fmt/290	AMI Draw Drawing		ami-draw		

Supported formats

ICC color profile	(ICC.1:2004-10)
JPEG 2000	JP2 (ISO/IEC 15444-1), JPX (ISO/IEC 15444-2)
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	TIFF 4 – 6, Class B, F, G, P, R, Y, TIFF/EP (ISO 12234-2), TIFF/IT (ISO 12639), GeoTIFF, Exif (JEITA CP-3451), DNG
UTF-8	ASCII (ANSI X3.4)
WAVE	BWF (EBU N22-1997)
XML	
Zip	

Supported formats

netCDF

http://www.unidata.ucar.edu/software/netcdf

Grib

http://www.wmo.int/pages/prog/www/WDM/Guides/Guide-binary-2.html

- Developed by the Wegener Institute (Germany)
  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
  http://java.sun.com/javase/6/docs/technotes/guides/language/annotations.html
- 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

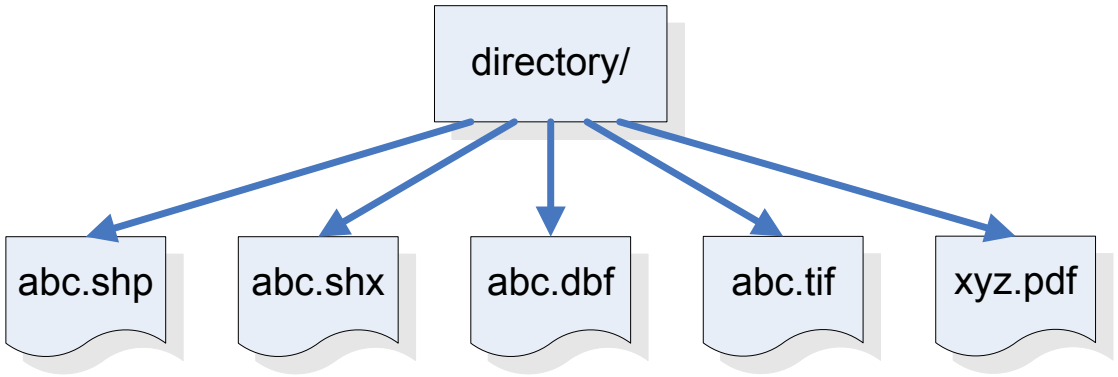
- Unitary
  - ✓ File — e.g. TIFF
  - ✓ File inside of a container — e.g. TIFF inside a Zip
  - ✓ Byte stream inside a file — e.g. ICC inside a TIFF
- Aggregate
  - ✓ Directory
  - ✓ Directory inside of a container
  - ✓ File set — e.g. command line arguments
  - ✓ Clump — e.g. Shapefile

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

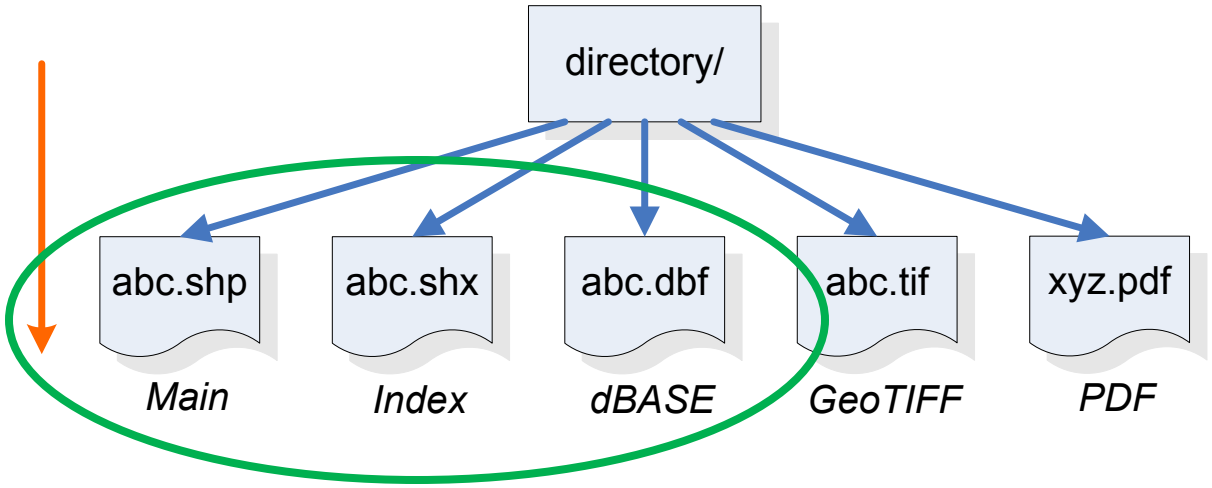
Characterization strategy

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*)

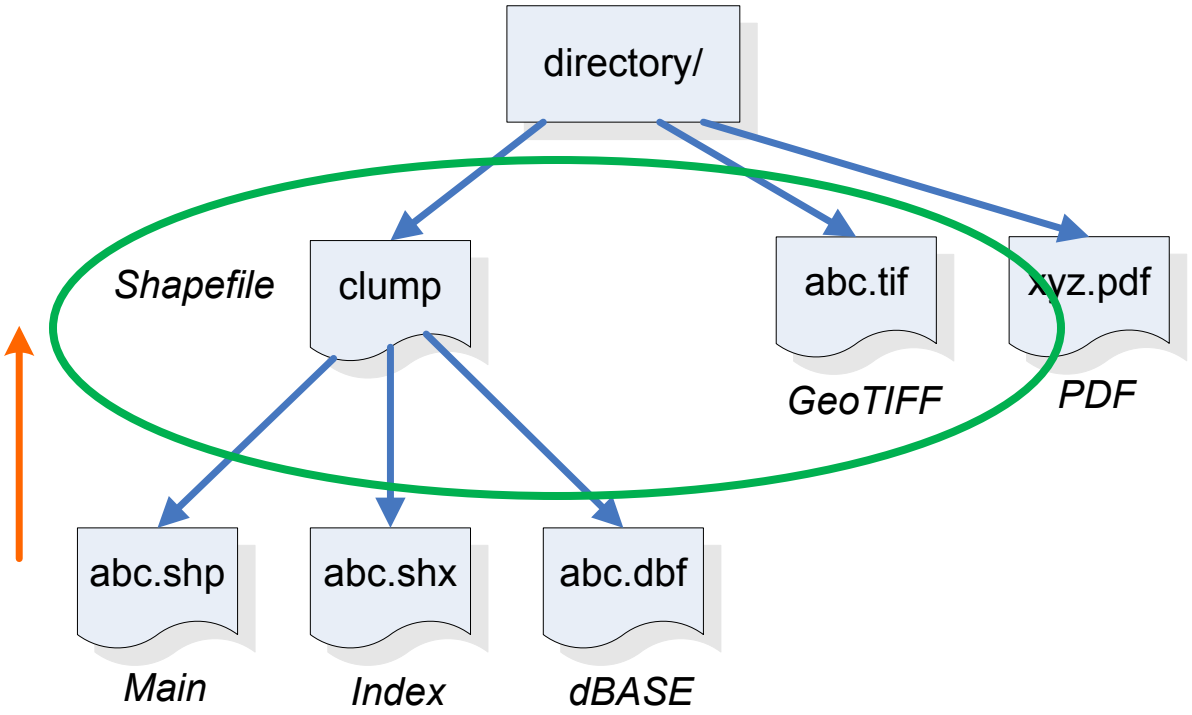
Characterization strategy



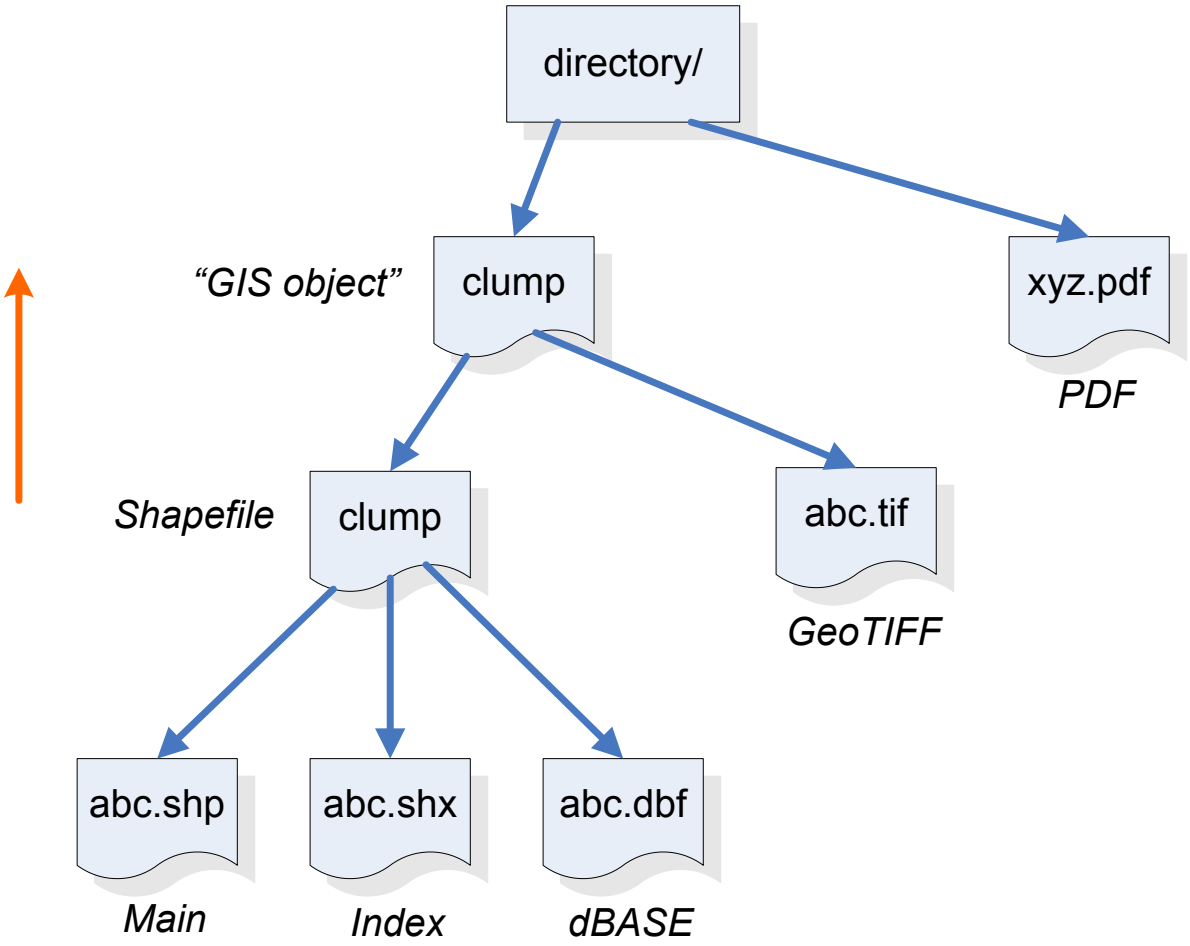
Characterization strategy



Characterization strategy



Characterization strategy



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
  - ∀ "for all"
  - ∃ "there exists" or "for any"

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

  property relation value

- Supported relational operators

  == != < > =< => 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/

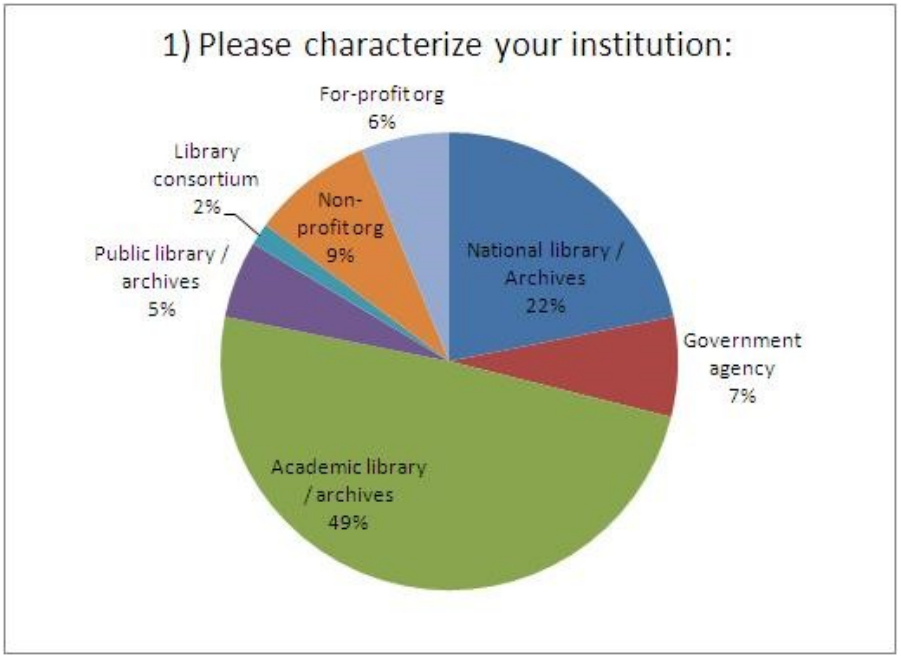
Demonstration

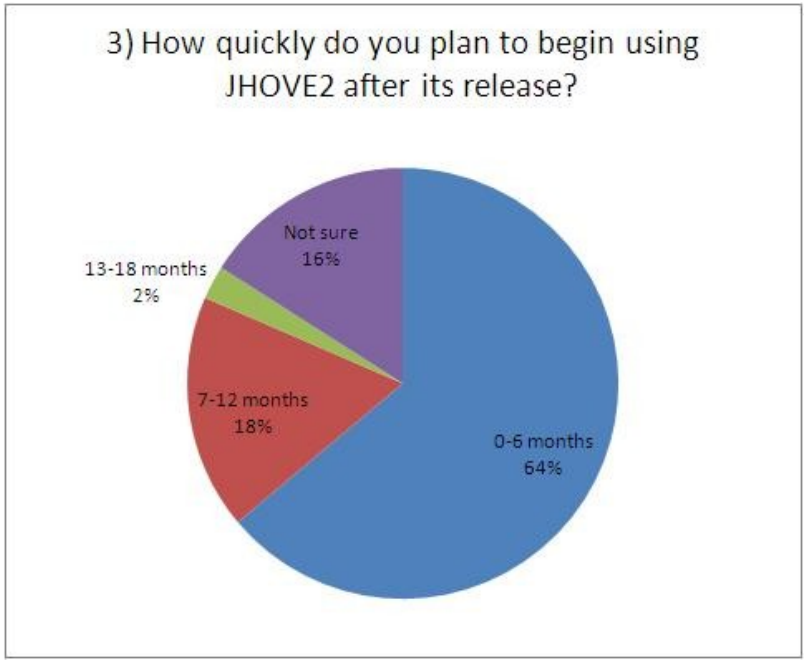
```
% jhove2 [-ik] [-b size]
        [-B Direct|NonDirect|Mapped]
        [-d JSON|Text|XML] [-f limit]
        [-t temp] [-o file] file ...
```

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

User survey

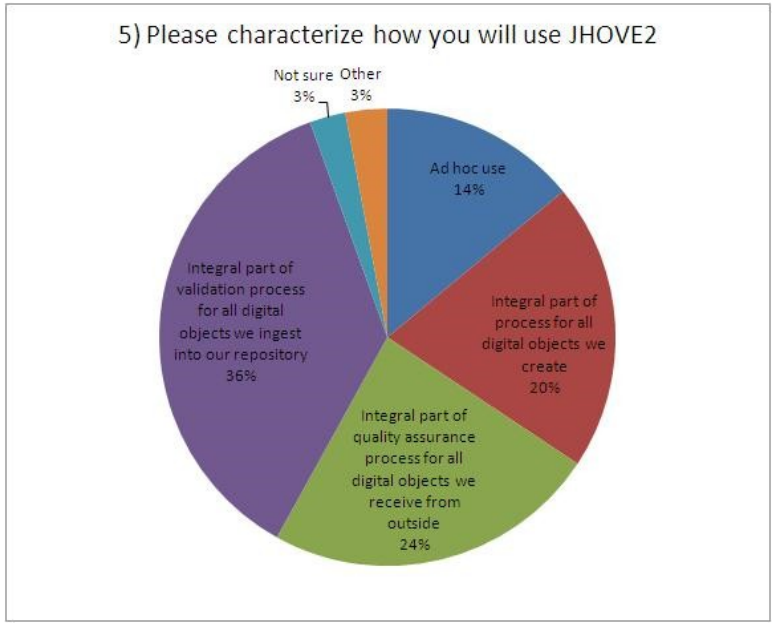
145 respondents, 88 institutions, 23 countries

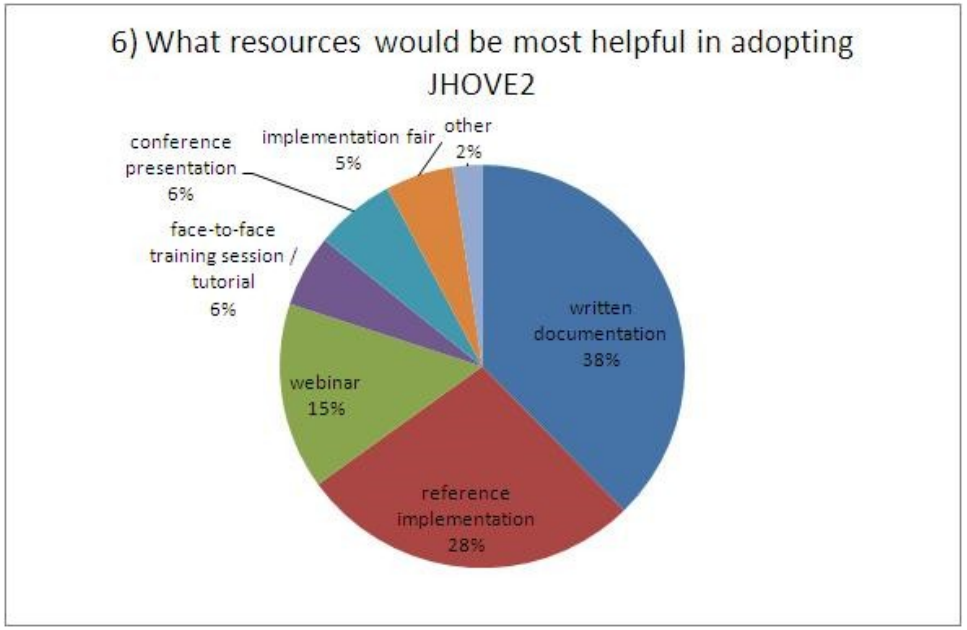




Full results available at https://confluence.ucop.edu/display/JHOVE2Info/User+survey

User survey





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

JHOVE2

UC Curation Center CDL California Digital Library

PORTICO

Stanford University LIBRARIES & ACADEMIC INFORMATION RESOURCES

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

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