

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*

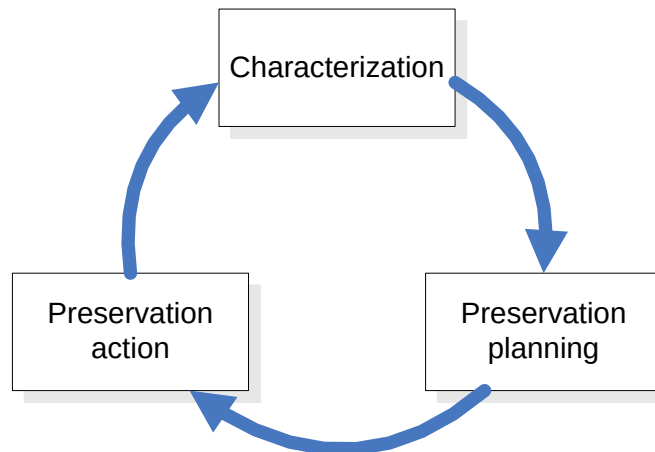
Day 2 agenda

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

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)?

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

“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
<http://www.planets-project.eu/market-survey/reports/>
- 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)

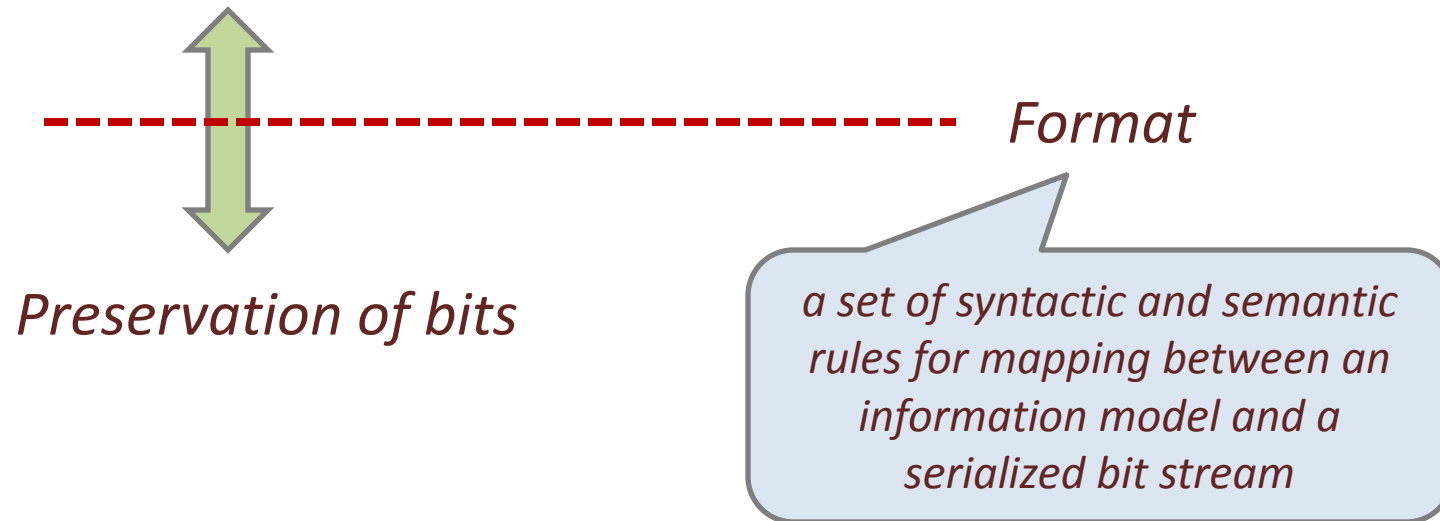
Characterization

- 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



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

Characterization

- Without strong format typing, all content is opaque

ffd8ffe000104a46494600010201
008300830000ffed0fb050686f74
6f73686f7020332e30003842494d
03e90a5072696e7420496e666f00
000000780000000000004800480000
000002f40240ffeeffee03060252
0347052803fc0002000000480048
0000000002d80228000100000064
0000000100030303000000001270f
0001000100000000000000000000
0000600800190190000000000000
0000000000000000000000000000
000000000000000000000000003842
494d03ed0a5265736f6c7574696f
6e00000000010008313a3000200 ...

SOI
APP0 JFIF 1.2
APP13 IPTC
APP2 ICC
DQT
SOF0 183x512
DRI
DHT
SOS
ECS0
RST0
ECS1
RST1
ECS2
...



Format characterization

- The automated determination of the intrinsic and extrinsic properties of a formatted object

- Identification
- Feature extraction
- Validation
- Assessment

Objective

Subjective

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

ff d8 ffe0 00 10 4a 46 49 46 00 01 02 01 00 83 00 83 00 00 ff ed 0f b0 50 68 6f ...

significant

"This JPEG is at the level of conformance to the normative requirements of a format's authoritative specification"

"This JPEG does not have a required 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

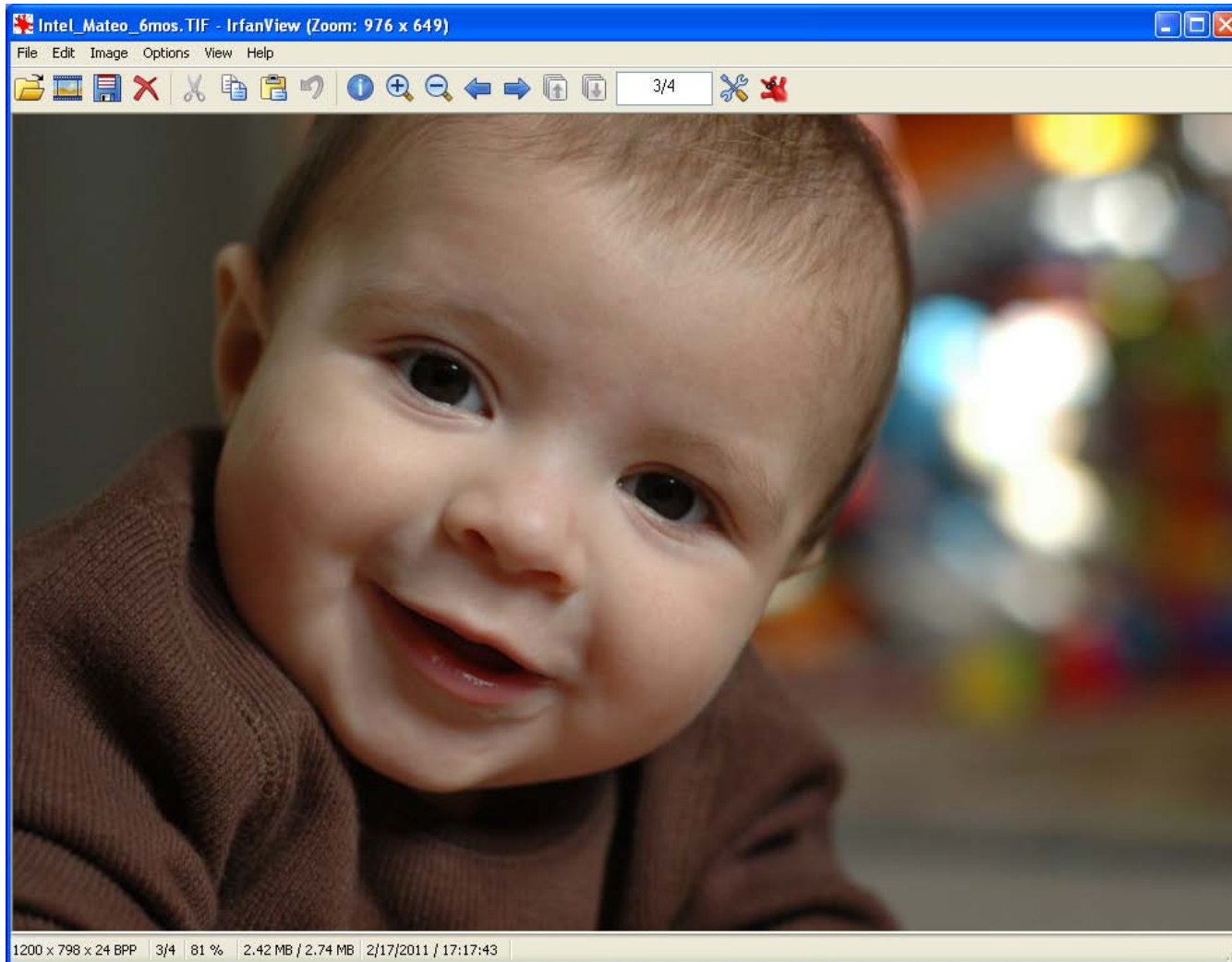
Format profiles

- 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

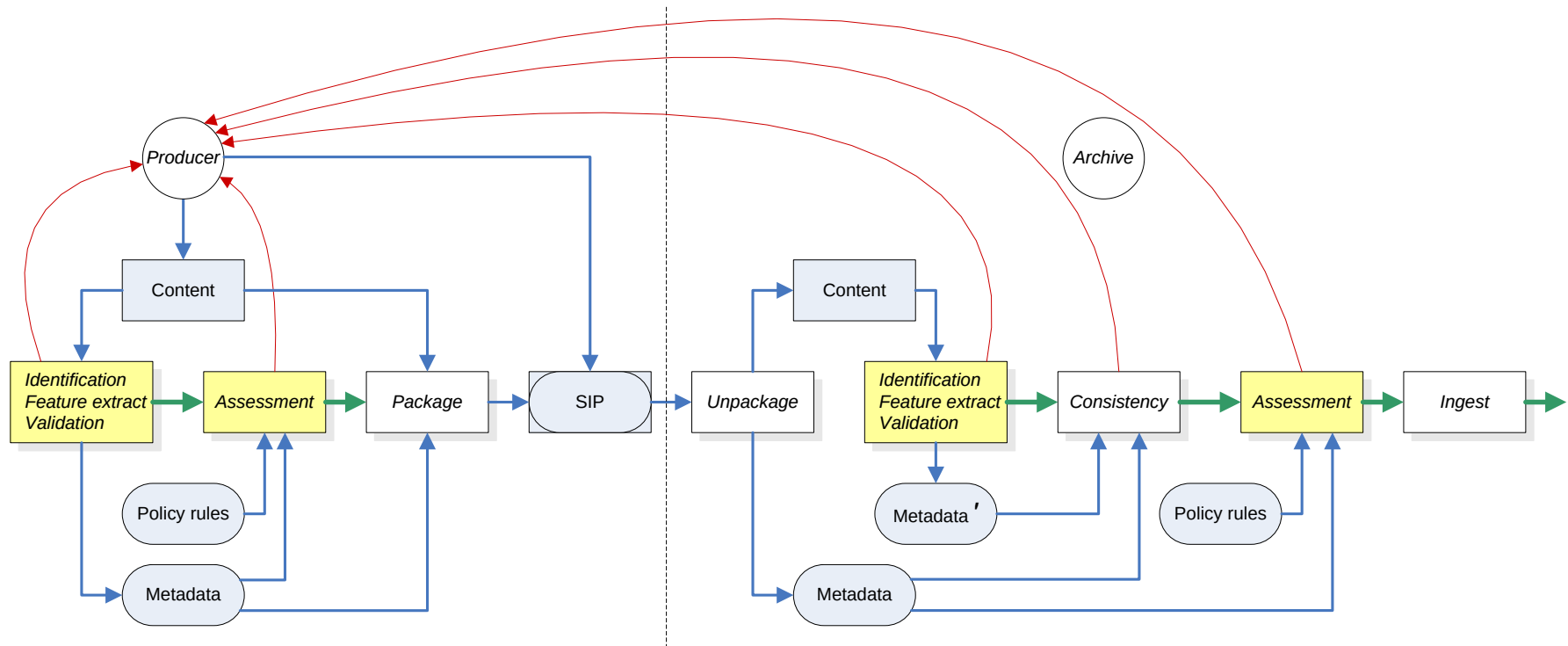
Characterization tools

- 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

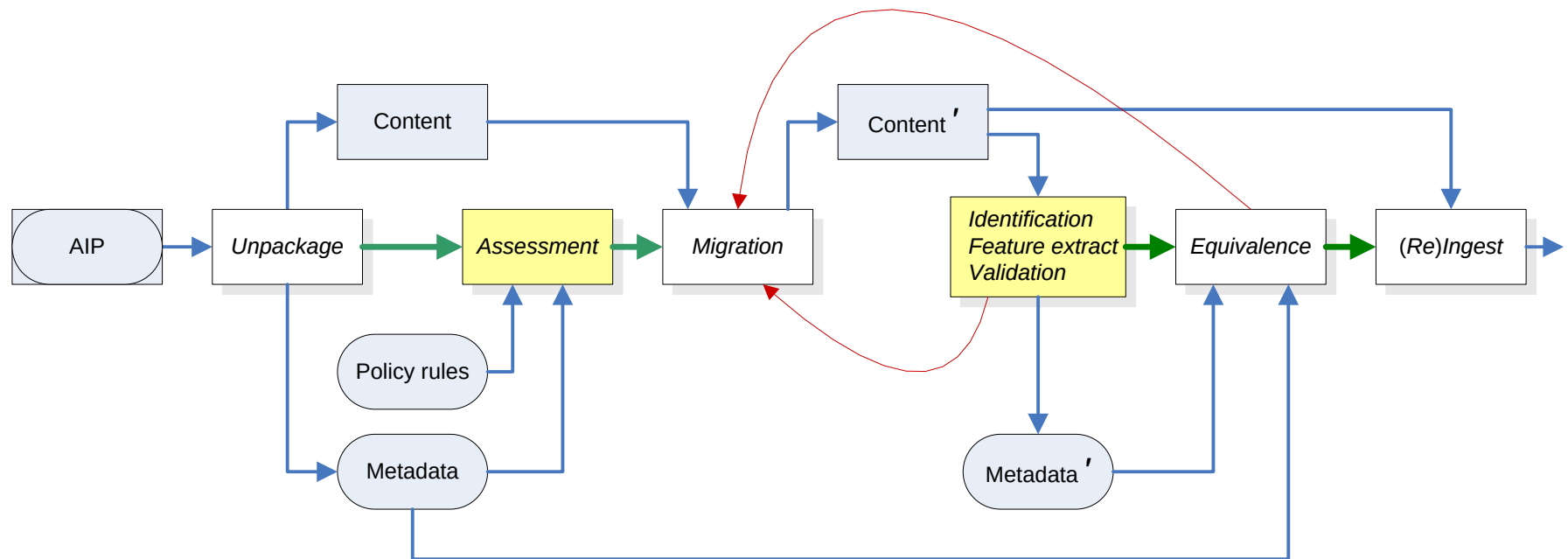
Demonstration



Characterization in ingest workflows



Characterization in migration workflows



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?



Curious Oysters, <http://www.flickr.com/photos/thecuriosoysters/4458657148/>

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*

Day 3 agenda

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

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

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

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

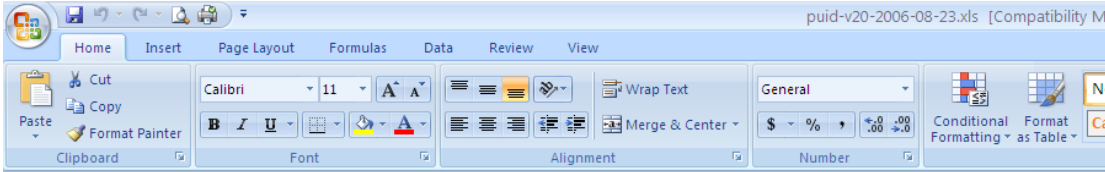
- Provide enhancements to JHOVE2 functionality
 - Extensive configuration via Spring dependency injection and Java properties files
 - Characterization strategy
 - Module customization
 - Message localization
- Complete documentation
 - User's guide
 - Module specifications
 - Architectural overview
 - 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
<http://www.nationalarchives.gov.uk/PRONOM>



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

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, ...
- TIFF 4 – 6, Class B, F, G, P, R, Y, TIFF/EP, TIFF/IT, Exif, GeoTIFF, DNG
- 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

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

Bibliothèque nationale de France / Atos Origin

- Gzip

<http://www.gzip.org/zlib/rfc-gzip.html>

- ARC

<http://www.archive.org/web/researcher/ArcFileFormat.php>

(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

- <http://java.sun.com/javase/6/docs/technotes/guides/language/annotations.html>

- 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)

- <http://www.oracle.com/database/berkeley-db/je/index.html>

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

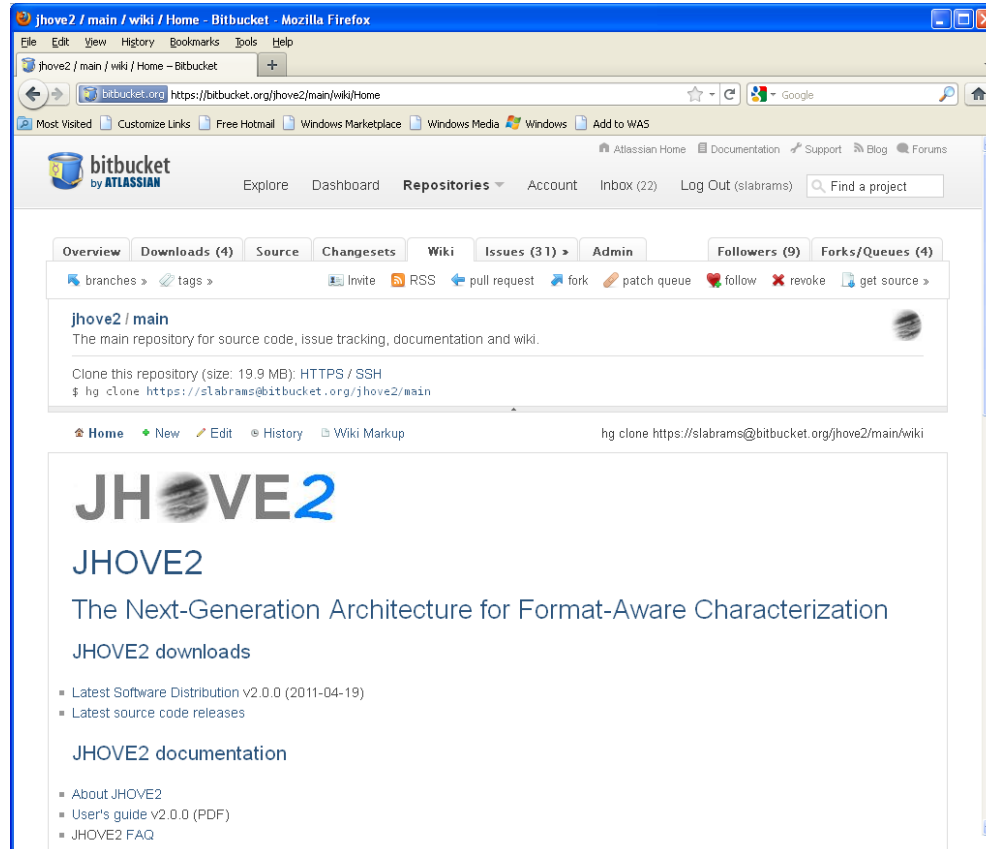
- *Bibliothèque nationale de France*
- *Deutsche Nationalbibliothek*
- *Ex Libris*
- *Fedora Commons / Rutgers University*
- *Florida Center for Library Automation*
- *Harvard University / GDFR project*
- *Koninklijke Bibliotheek*
- *Library of Congress*
- *MIT / DSpace*
- *NARA*
- *National Library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

JHOVE2 project

<http://jhove2.org/>

JHOVE2-Announce-L@listserv.ucop.edu

JHOVE2-Techtalk-L@listserv.ucop.edu



The screenshot shows the Bitbucket web interface for the 'jhove2 / main' repository. The browser window title is 'jhove2 / main / wiki / Home - Bitbucket - Mozilla Firefox'. The address bar shows the URL 'https://bitbucket.org/jhove2/main/wiki/Home'. The Bitbucket logo and navigation links (Explore, Dashboard, Repositories, Account, Inbox (22), Log Out (slabrams)) are visible. The repository page includes tabs for Overview, Downloads (4), Source, Changesets, Wiki, Issues (31), Admin, Followers (9), and Forks/Queues (4). The 'Overview' tab is selected, showing the repository name 'jhove2 / main', a description 'The main repository for source code, issue tracking, documentation and wiki.', and clone instructions: 'Clone this repository (size: 19.9 MB): HTTPS / SSH' and '\$ hg clone https://slabrams@bitbucket.org/jhove2/main'. Below this, there are links for Home, New, Edit, History, and Wiki Markup. The main content area features the JHOVE2 logo and title, followed by the subtitle 'The Next-Generation Architecture for Format-Aware Characterization'. Under 'JHOVE2 downloads', there are links for 'Latest Software Distribution v2.0.0 (2011-04-19)' and 'Latest source code releases'. Under 'JHOVE2 documentation', there are links for 'About JHOVE2', 'User's guide v2.0.0 (PDF)', and 'JHOVE2 FAQ'.

Day 3 agenda

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

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/JH0VE2>

- Property and message identifiers are based on the underlying class name and accessor method

<http://jhove2.org/terms/property/org/jhove2/core/JH0VE2/Commands>

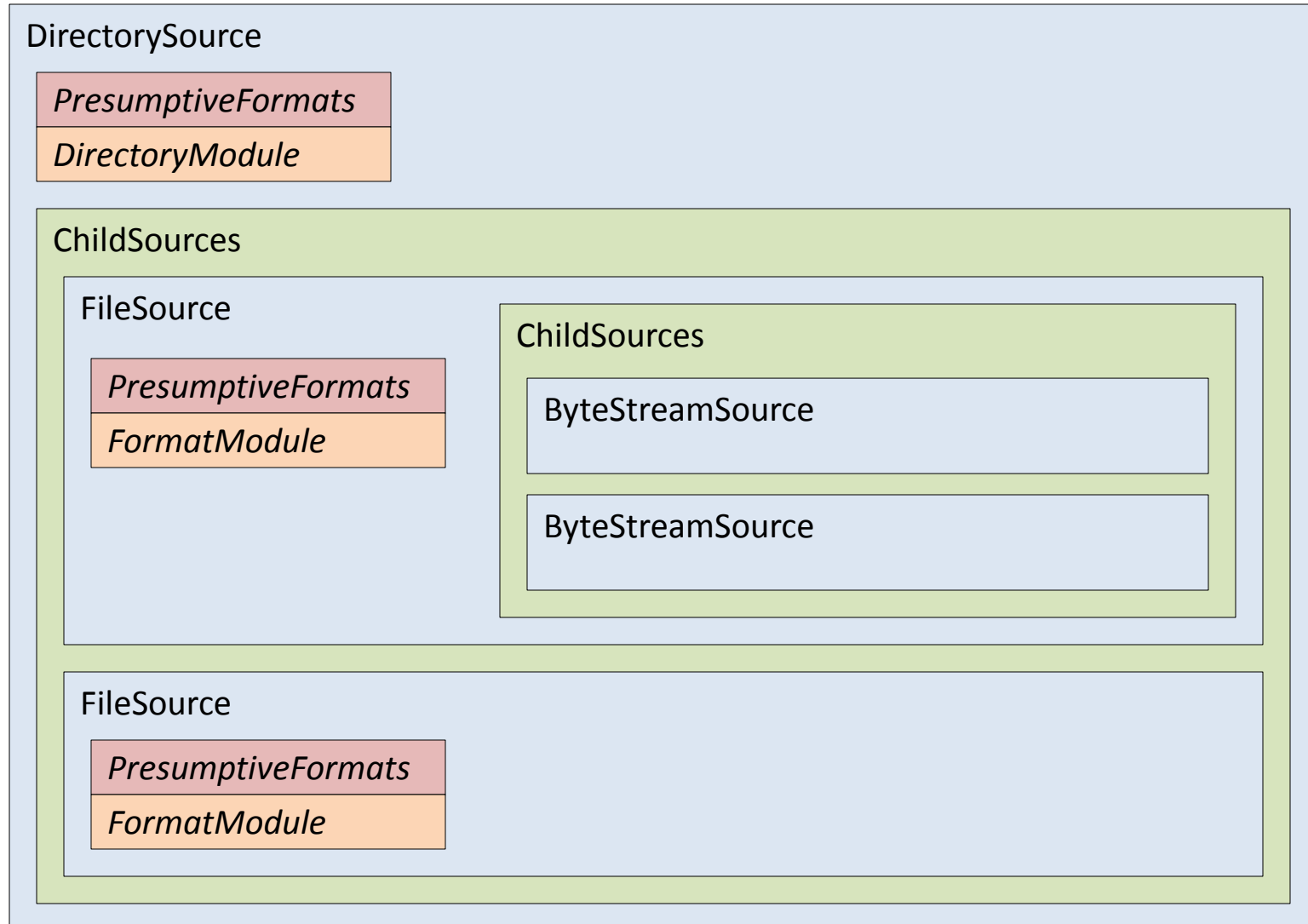
<http://jhove2.org/terms/message/org/jhove2/module/utf8/UTF8Module/ByteOrderMark>

- 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

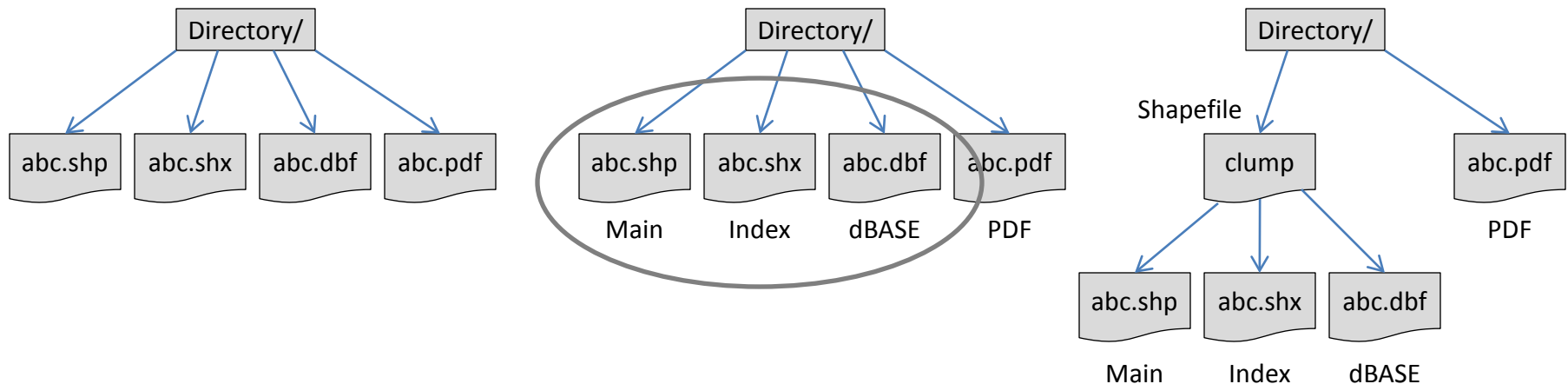


Source unit

- 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

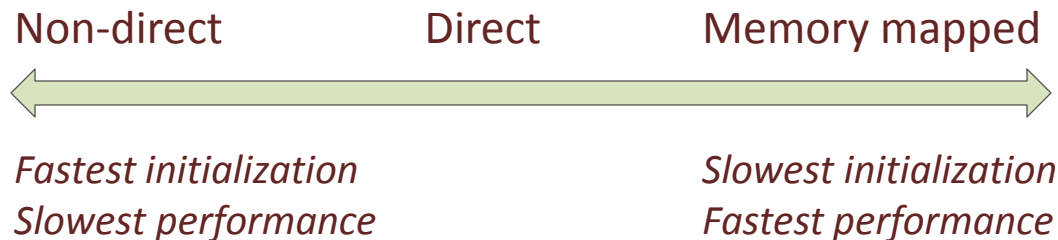
- A source unit may be associated with messages documenting various conditions
 - Error *A terminal condition requiring remedial action*
 - Warning *A condition possibly requiring attention*
 - Informative *A condition not requiring further attention*
- Messages can arise in two contexts
 - Process *An unanticipated condition arising from the characterization process*
 - Object *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 *I/O subsystem of underlying OS*
 - Non-direct *I/O subsystem of JVM*
 - Memory mapped *Paging subsystem of underlying OS (1.6 GB limit)*



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

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

Day 3 agenda

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

Command line invocation

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

-i	--show-identifiers
-k	--calc-digests
-b size	--buffer-size size
-B type	--buffer-type type
-d <i>displayer</i>	--display <i>displayer</i>
-t temp	--temp temp
-o out	--output out
<i>file</i>	

Show identifiers in JSON and Text displays

Calculate message digests

I/O buffer size, in octets (*default: 131072*)

I/O buffer type: Direct, NonDirect, Mapped

Display: JSON, Text, XML

Temporary directory (*default: java.io.tmp*)

Output file (*default: standard output*)

File, directory, or URL

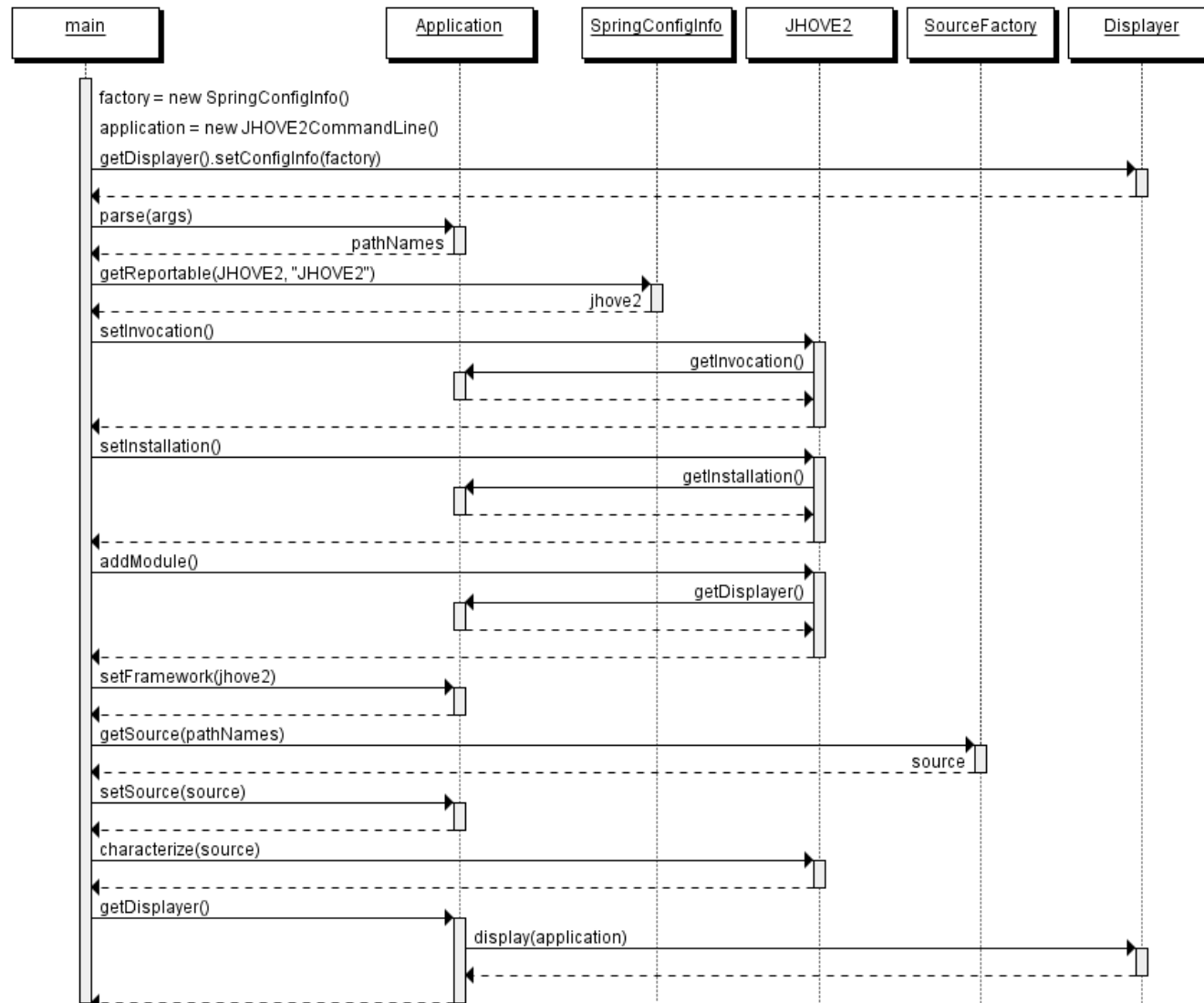
Day 3 agenda

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

Architecture

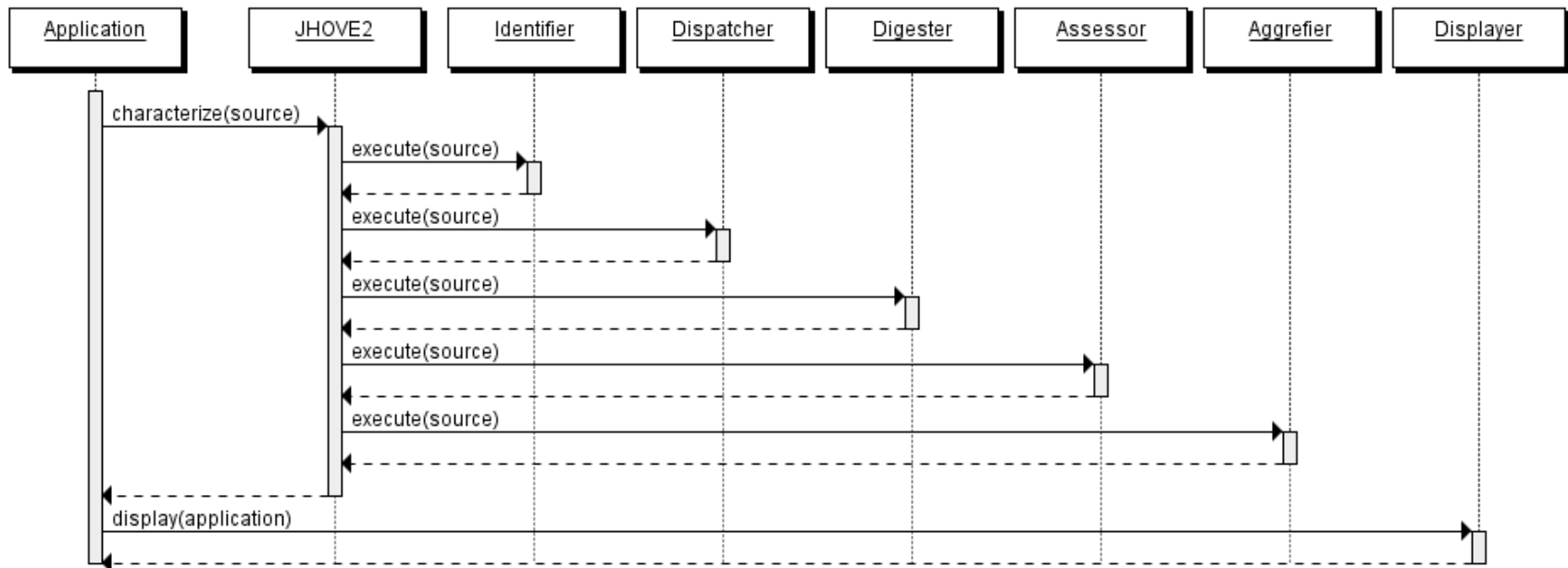
- Three conceptual layers
 - Application *Command-line application*
 - Framework *Coordinates all processing*
 - Modules *Embodies specific processing behaviors*
- Technically, the JHOVE2 application and framework are implemented as modules, however, due to their central, but distinct, roles in processing, it is useful to consider them as conceptually independent levels

Application



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
 - AggregfierCommand $\Rightarrow$ AggregfierModule
 - Pathname globbing identification

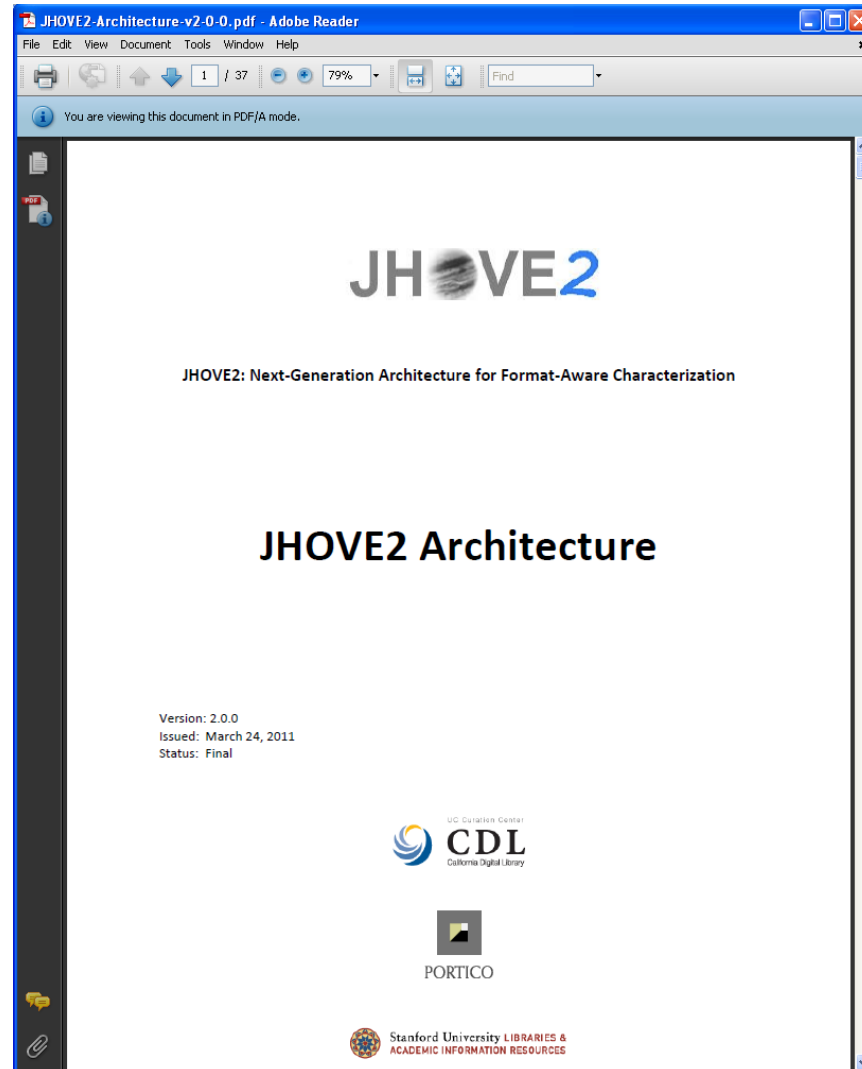
Format modules

- 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

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

Day 3 agenda

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

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

jhove2 / main / downloads - Bitbucket - Mozilla Firefox

File Edit View History Bookmarks Tools Help

bitbucket.org https://bitbucket.org/jhove2/main/downloads

Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Add to WAS

bitbucket by ATlassian

Explore Dashboard **Repositories** Account Inbox (22) Log Out (slabrams) Find a project

Overview Downloads (4) Source Changesets Wiki Issues (31) Admin Followers (9) Forks/Queues (4)

branches » tags » Invite RSS pull request fork patch queue follow revoke get source »

jhove2 / main

The main repository for source code, issue tracking, documentation and wiki.

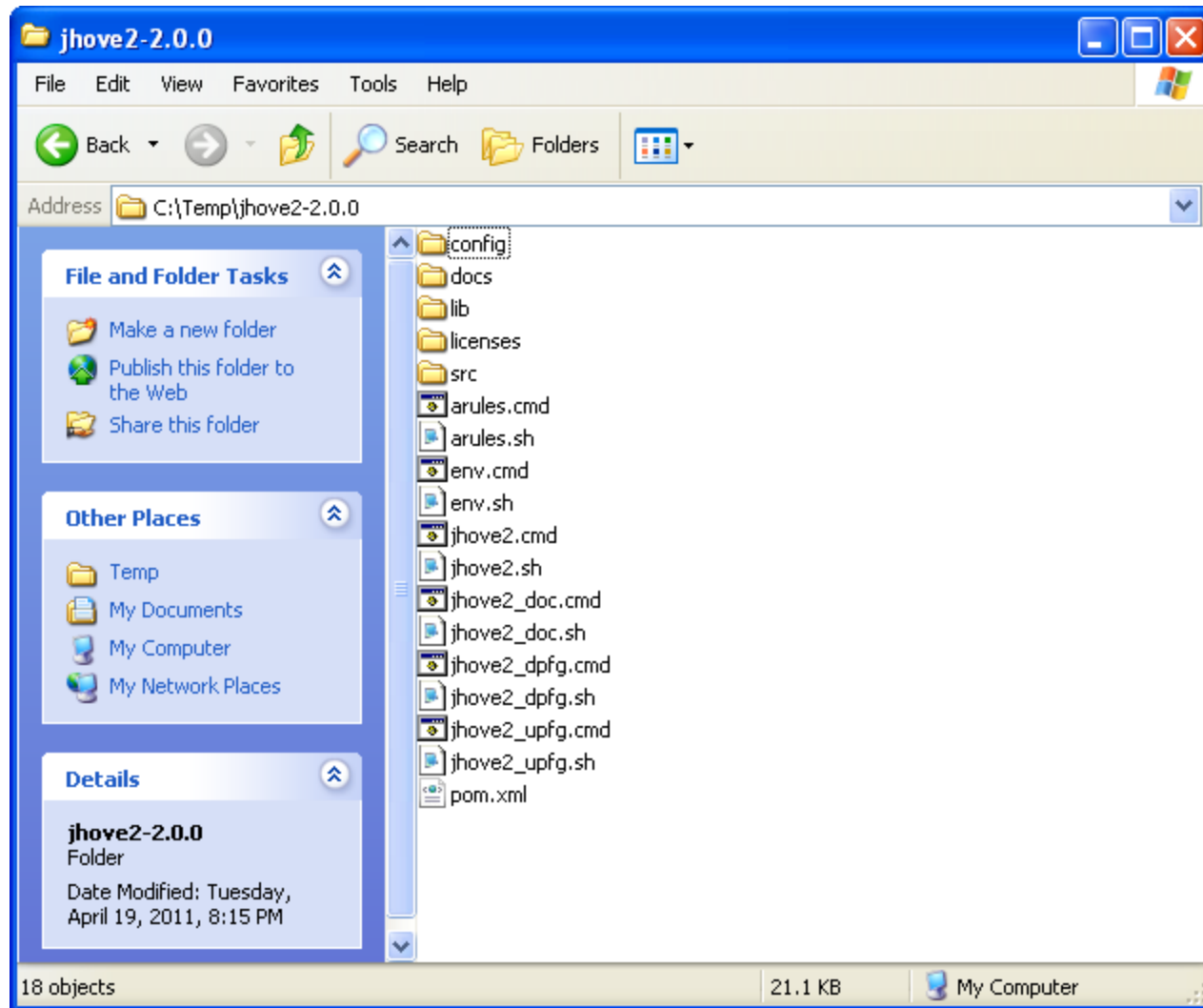
Clone this repository (size: 19.9 MB): HTTPS / SSH
\$ hg clone https://slabrams@bitbucket.org/jhove2/main

Uploaded files

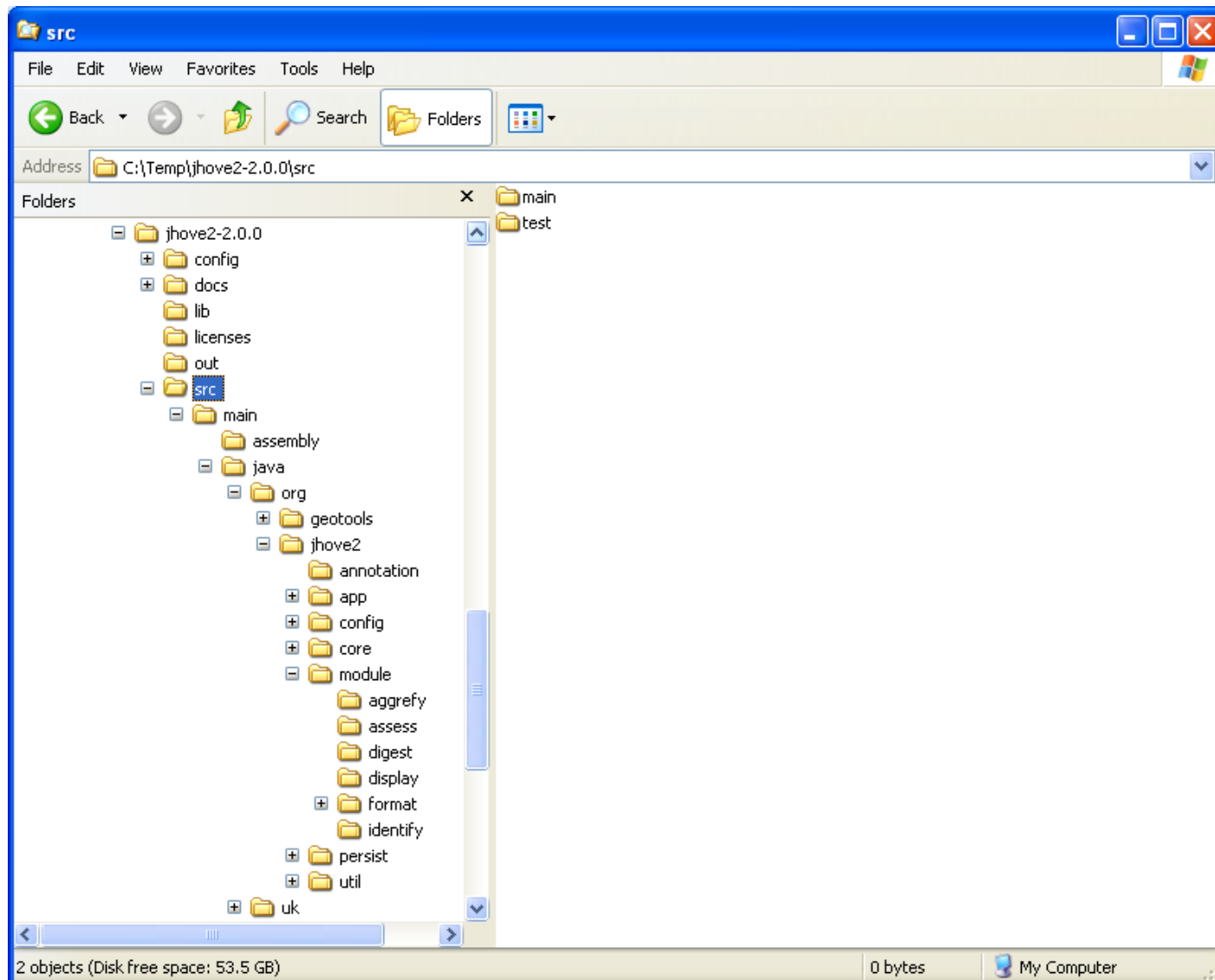
Filename	Size	Uploaded	Uploader	Downloaded
JHove2-2.0.0-release-notes.txt	6.2 KB	8 days ago	mstrong	10 times
JHove2-2.0.0-release-notes.pdf	499.9 KB	8 days ago	mstrong	14 times
jhove2-2.0.0.tar.gz	27.2 MB	10 days ago	mstrong	17 times
jhove2-2.0.0.zip	30.0 MB	10 days ago	mstrong	48 times

<https://bitbucket.org/jhove2/main/downloads>

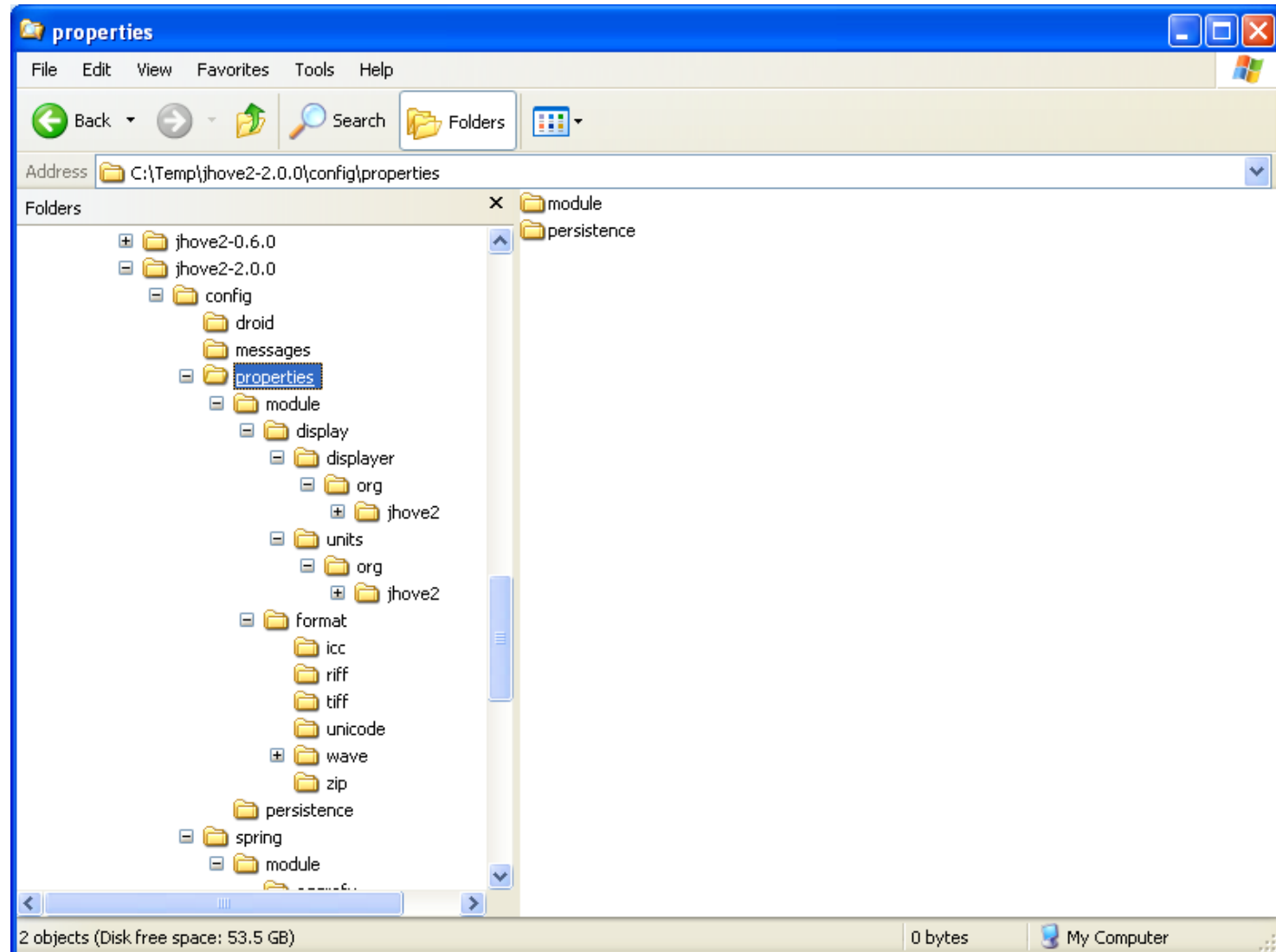
Disaggregate the Zip or tar.gz



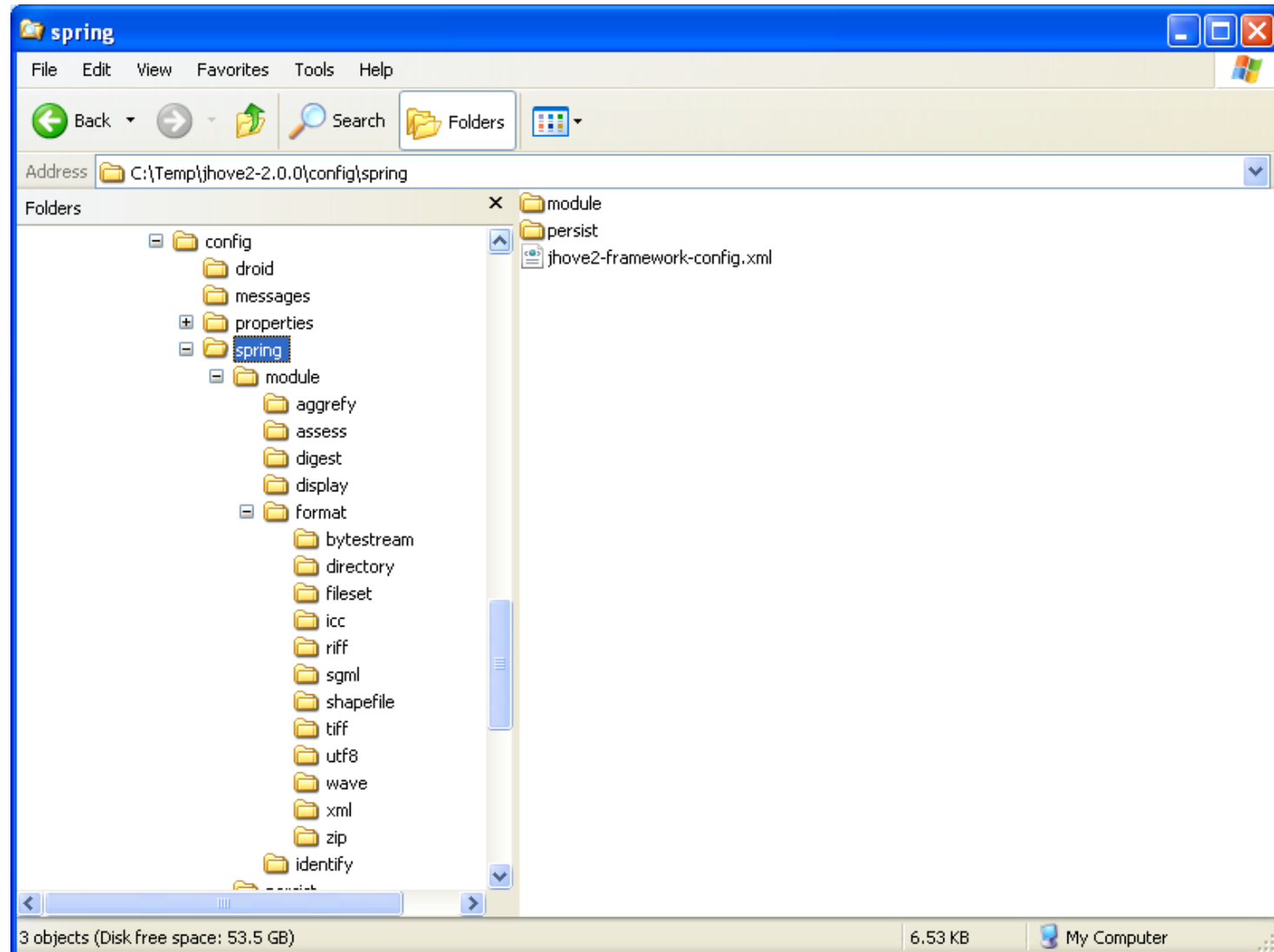
Installation directory structure



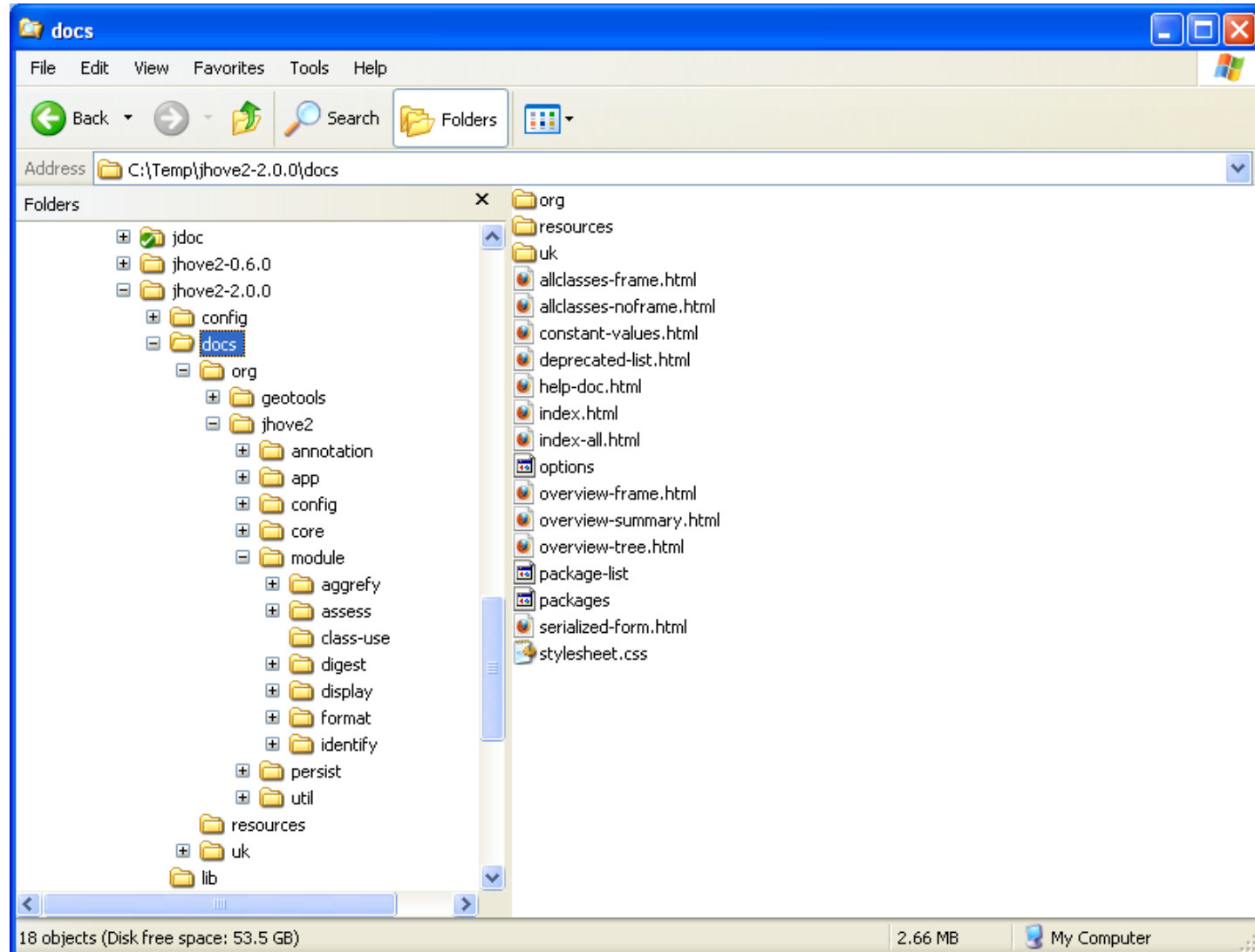
Installation directory structure



Installation directory structure



Installation directory structure



Day 3 agenda

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

Invocation scripts

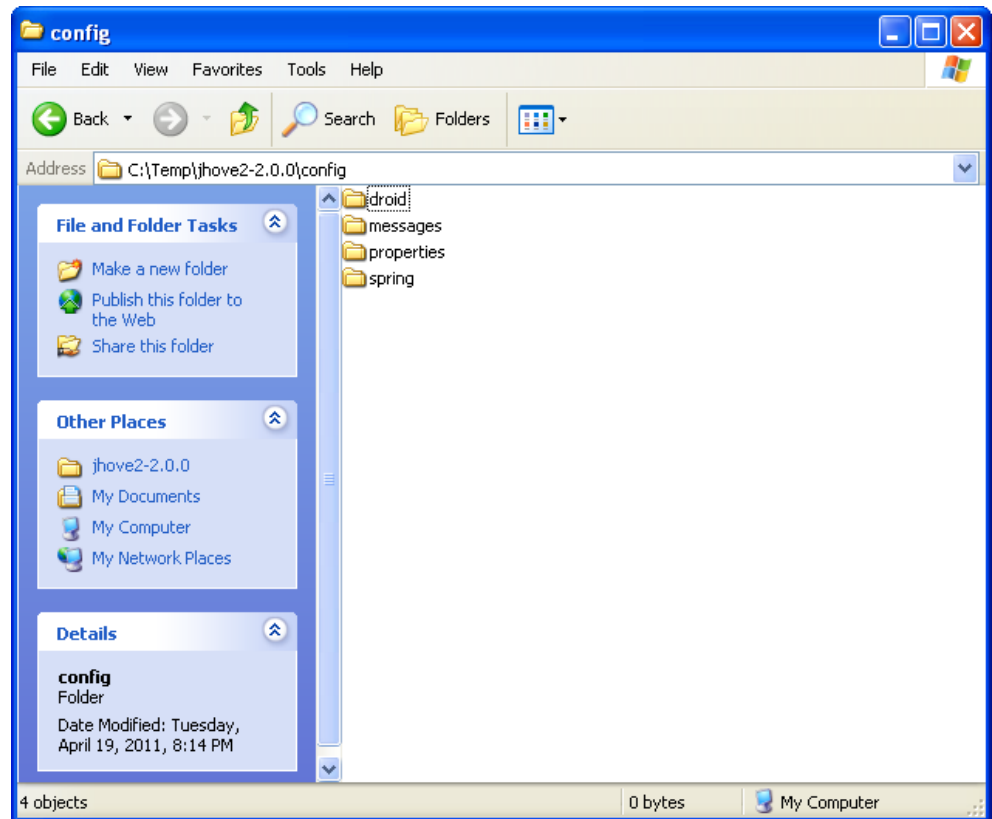
- All scripts are paired, one each for Unix (.sh) and Windows (.cmd)
 - arules.cmd *Assessment rule generator*
arules.sh
 - env.cmd *Java environment definition*
env.sh
 - jhove2.cmd *JHOVE2 command line application*
jhove2.sh
 - jhove2_doc.cmd *Module documentation generator*
jhove2_doc.sh
 - jhove2_dpfg.cmd *Displayer configuration generator*
jhove2_dpfg.sh
 - jhove2_upfg.cmd *Units of measure configuration generator*
jhove2_upfg.sh

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](#)

Display 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

config/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 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="AggregfierCommand"/>
    </list>
  </property>
  ...
</bean>
<bean id="IdentifierCommand" class="org.jhove2.module.identify.Identifier ... >
  ...
```

Spring configuration

- Digester algorithms

config/spring/module/digest/jhove2-digest-config.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

```
<!-- 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>
```

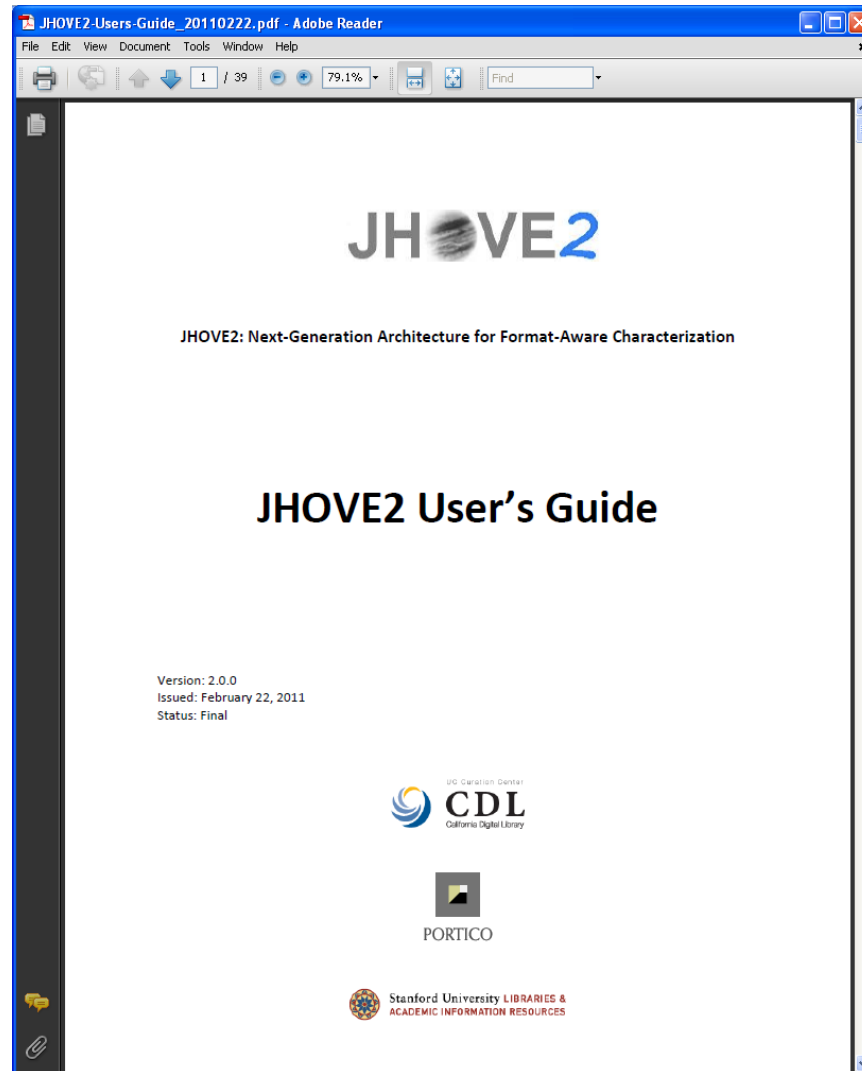
Spring configuration

- Persistence manager

`properties/persistence/persistence.properties`

```
# classname org.jhove2.config.spring.SpringInMemoryPersistenceManagerFactory  
classname org.jhove2.config.spring.SpringBerkeleyDbPersistenceManagagerFactory
```

User's guide



Day 3 agenda

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

Day 3 agenda

	<i>Time</i>	<i>Topic</i>
✓	08:30 – 08:35 am	<i>Review of objectives and agenda</i>
✓	08:35 – 09:15 am	<i>JHOVE2 project</i>
✓	09:15 – 10:15 am	<i>Concepts</i>
✓	10:15 – 10:45 am	<i>Break</i>
✓	10:45 – 11:30 am	<i>Demonstration</i>
✓	11:430– 12:00 pm	<i>Architecture</i>
✓	12:-0 – 13:30 pm	<i>Lunch</i>
✓	13:30 – 14:00 pm	<i>Installation</i>
✓	14:00 – 15:00 pm	<i>Configuration</i>
✓	15:00 – 15:30 pm	<i>Break</i>
	15:30 – 16:30 pm	Community building and sustainability
	16:45 – 17:00 pm	Questions and discussion

Day 3 agenda

	<i>Time</i>	<i>Topic</i>
✓	08:30 – 08:35 am	<i>Review of objectives and agenda</i>
✓	08:35 – 09:15 am	<i>JHOVE2 project</i>
✓	09:15 – 10:15 am	<i>Concepts</i>
✓	10:15 – 10:45 am	<i>Break</i>
✓	10:45 – 11:30 am	<i>Demonstration</i>
✓	11:30 – 12:00 pm	<i>Architecture</i>
✓	12:00 – 13:30 pm	<i>Lunch</i>
✓	13:30 – 14:00 pm	<i>Installation</i>
✓	14:00 – 15:00 pm	<i>Configuration</i>
✓	15:00 – 15:30 pm	<i>Break</i>
✓	15:30 – 16:30 pm	<i>Assessment</i>
	16:45 – 17:00 pm	Questions and discussion

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

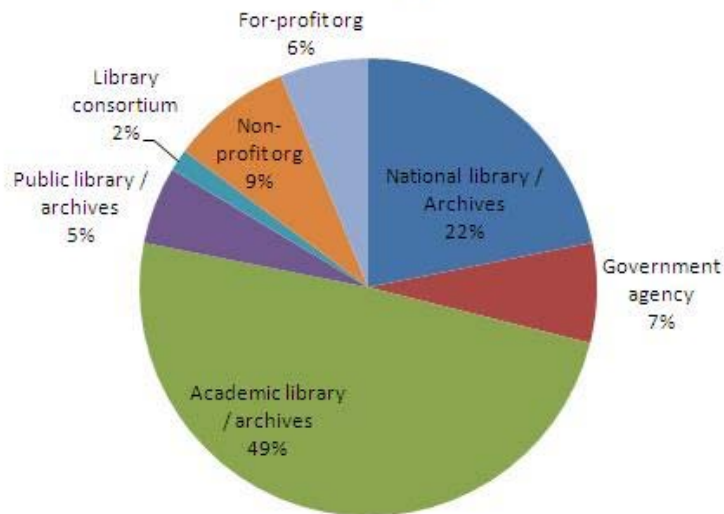
Community

- 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

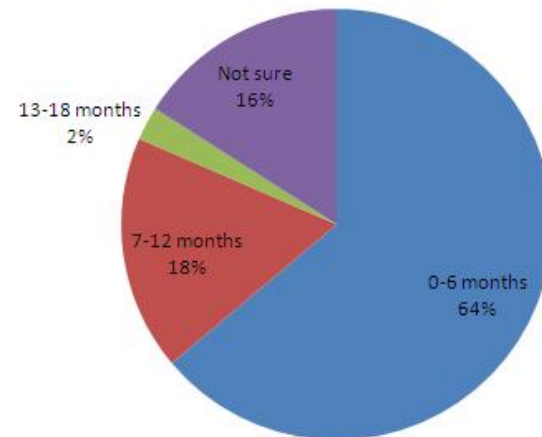
User-driven priorities

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

1) Please characterize your institution:



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

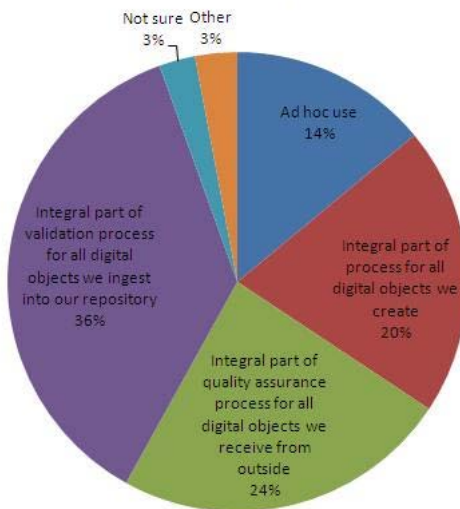


Full results available at https://bitbucket.org/jhove2/main/wiki/User_survey

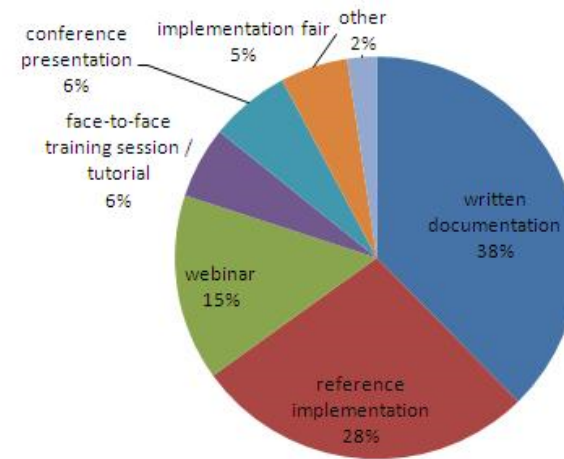
User-driven priorities

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

5) Please characterize how you will use JHOVE2



6) What resources would be most helpful in adopting JHOVE2



Full results available at https://bitbucket.org/jhove2/main/wiki/User_survey

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?



Curious Oysters, <http://www.flickr.com/photos/thecuriosoysters/4458657148/>

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*