Open Preservation Foundation

JHOVE

Evaluation & Stabilisation Plan

Open Preservation Foundation

March 2015

Introduction

This report details the Open Preservation Foundation's (OPF) plans to provide a permanent, sustainable home for the JHOVE project.

JHOVE started as a joint development effort by JSTOR and Harvard University Library, partly funded by the Andrew W. Mellon Foundation. The source code and documentation was moved to SourceForge in March 2008 by Gary McGath. The SourceForge project is owned by Gary who is also the named maintainer. Gary oversaw 11 JHOVE releases on SourceForge, the last of which was JHOVE version 1.11 in September 2013. Gary then moved the project to GitHub, GitHub having overtaken SourceForge as the open source project hosting platform of choice.

In March 2014 Gary announced that he was no longer willing to continue as sole maintainer of JHOVE. The OPF have agreed to take over stewardship of JHOVE which was highlighted in their technology as one of the most widely used pieces of software by members and the wider community.

This document describes the OPF's initial activities and aims. We list all of the JHOVE resources that we intend to maintain or preserve. Each resource is described while examining it's history and current level of use. We include some items that are external to the JHOVE project such as authoritative how tos.

The final section describes the OPF's immediate plans for taking ownership of and responsibility for the JHOVE project assets and our progress so far.

Existing JHOVE Resources

This section catalogues the current state of JHOVE online resources. The move to OPF stewardship provides an opportunity to rationalise and update the JHOVE materials.

SourceForge Project

Hosted at http://sourceforge.net/projects/jhove/, the project was last updated 18 months ago on 30/09/2013. The project license is stated as GNU Library or Lesser General Public License version 2.0. All downloadable archives state their license as GNU Library or Lesser General Public License version 2.1, rather than 2.0. The project has thirteen 5 star reviews, the last was written at the time of the latest 1.11 release in September 2013.

Source Code & History

The project's source is hosted in a CVS repository, https://sourceforge.net/p/jhove/cvs/. The revision history can be browsed online here. Source code history dates back to 23rd March 2008 when the project was added to Sourceforge. The initial commit makes reference to previous history at HUL JHOVE repository but doesn't include a link. We will follow up with with Gary McGath, previous maintainer of JHOVE, to see if it can be found anywhere.

Documentation

The JHOVE website is hosted through SourceForge here: http://jhove.sourceforge.net/. The site has an organically grown feel to it and would benefit from restructuring and rewriting. The current structure is:

- **Home:** an overview of various JHOVE related material.
- **Tutorial:** a "using JHOVE" tutorial that covers everything from installation, through to GUID use.
- **Documentation:** A site index in the form of a structured set of links to some of the site pages and external resources.
- **Distribution:** Link to the download page on SourceForge plus installation requirements and instructions.
- **Links:** A few links to external resources.

All of these resources should be QA'd for accuracy and incorporated in the new JHOVE site.

Downloads

The JHOVE file download page gives users access to:

- a folder containing sub-folders, one for each JHOVE release; and
- examples.zip, a zip archive that contains example content sorted into folders by content type: ASCII, GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, & XML.

The release sub-folders are identified by sensible names, e.g. “JHOVE 1.1”, as shown in the screenshot below.

Home / jhove

Name	Modified	Size	Downloads / Week
Parent folder			
JHOVE 1.11	2013-09-30		33
JHOVE 1.11b1	2013-09-22		1
JHOVE 1.10	2013-06-10		3
JHOVE 1.9	2013-03-04		1
JHOVE 1.8	2012-11-07		1
JHOVE 1.7	2012-08-12		1
JHOVE 1.6	2011-01-04		1
JHOVE 1.5	2009-12-23		1
JHOVE 1.4	2009-07-31		1
JHOVE 1.3	2009-06-04		1
JHOVE 1.2	2009-02-11		1
JHOVE 1.1	2008-04-28		2

Totals: 12 Items

Early releases consisted of two archive files, a zip and a gzipped tar file, with identical contents. Over time standard extras have been added, a README, checksums, and release notes. These are also contained in the downloadable archives. A screenshot of the latest

1.11 release page is shown below:

Looking for the latest version? **Download jhove-1_11.tar.gz (4.0 MB)**

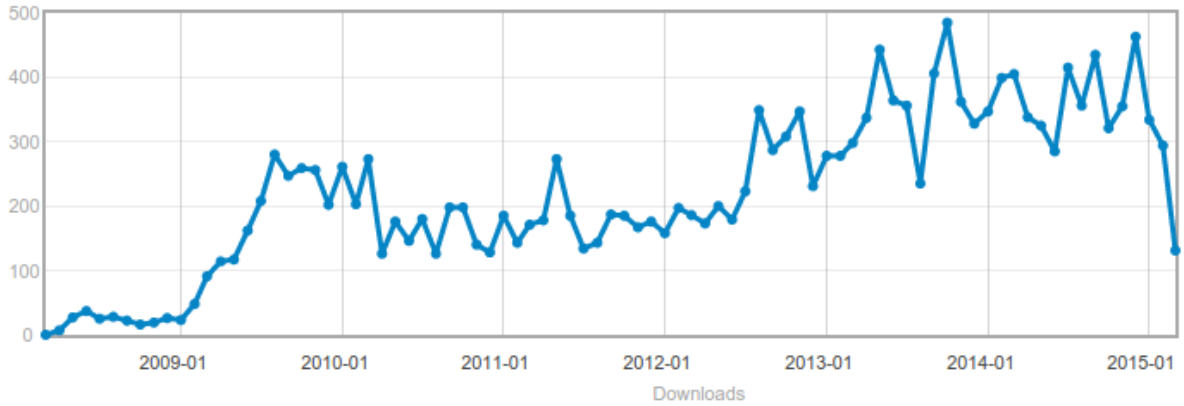
Home / jhove / JHOVE 1.11

Name	Modified	Size	Downloads / Week
Parent folder			
JHOVE 1.11.app.zip	2013-09-30	745.8 kB	1
RELEASENOTES	2013-09-30	62.5 kB	2
README	2013-09-30	10.6 kB	2
jhove-1_11.tar.gz	2013-09-30	4.0 MB	7
jhove-1_11.zip	2013-09-30	5.5 MB	26
Totals: 5 Items		**10.3 MB**	**38**

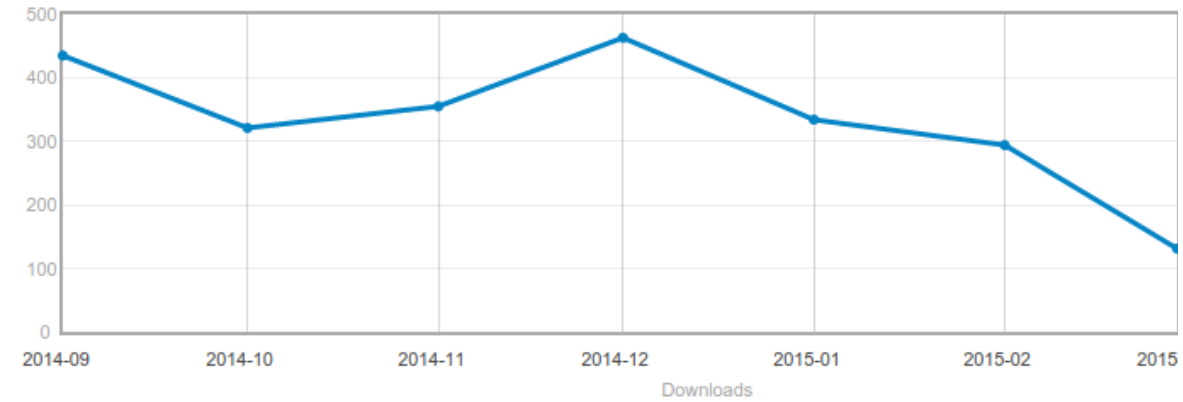
The downloads are not one click installs. The two archive formats simply reflect the popularity of zip on the windows platform, while gzipped tar archives are preferred by linux users.

There have been 300 - 400 downloads a week since early 2013 but this has tailed off significantly in recent months. This can be seen in the graphs below, the first is a complete download history from January 2008, the second shows downloads in the last 6 months.

JHOVE downloads since 01/01/2008



JHOVE downloads since 17/09/2014



The reason for the downward trend in activity isn't certain but the frequent releases made in 2009, 2012, and 2013 do seem to have helped drive downloads. The table below shows the number of JHOVE releases made by year, there haven't been any releases since September 2013.

Year	2008	2009	2010	2011	2012	2013
Releases	1	4	0	1	2	3

Issue Tracking

There's historical use of the SourceForge issue tracker, the most recent activity was 14th May 2014. The issues are divided into sub-categories and have assigned statuses as shown in the table below:

	Open	Closed	Last Activity
Bugs	15	45	14/05/2014
Support Requests	1	1	06/03/2013
Patches	2	1	12/07/2012
Feature Requests	6	2	17/04/2013

These will need to be triaged with the issues recorded in the current GitHub repository. The history will be retained as well as any unfixed issues.

Mailing List

There is a jhove-discuss mailing list that's hosted by SourceForge, activity is approximately a single post per month. The mailing list will be moved but the subscribers and archive would have to be preserved also.

Discussion Forum

There's a JHOVE discussion forum that was last used in 11th July 2014. There are only 12 discussion threads, the most read has received 301 views. One of the threads is SourceForge specific, "...where the sources???", the rest discuss software functionality and should be preserved.

Wiki

The wiki consists of a single page with a one line description of JHOVE and a link to the latest download.

GitHub Project

The JHOVE GitHub project can be found at Gary McGath's GitHub page, the repo URL is https://github.com/gmcgath/jhove. The LICENSE and COPYING files both state the license as GNU LGPL version 2.1. The project currently has 8 watchers and 10 GitHub "stargazers". There are currently 12 forks, though forks and pull requests are GitHub's standard mechanism for distributed development.

Source Code & History

The projects commit history dates back to 23rd March 2008, identical to the initial SourceForge commit. If the previous source history could be found it would be possible to add it to the project.

All JHOVE development activity since the 1.11 release on 30/09/2013 has used Git for revision control and GitHub for online hosting. The last development activity in the master branch was the addition of Unicode 7 support on 18th June 2014. The only activity since then has been an update to the README by Gary on 3rd February 2015. This marks the repository as archival and hands maintenance to the Open Preservation Foundation.

Documentation

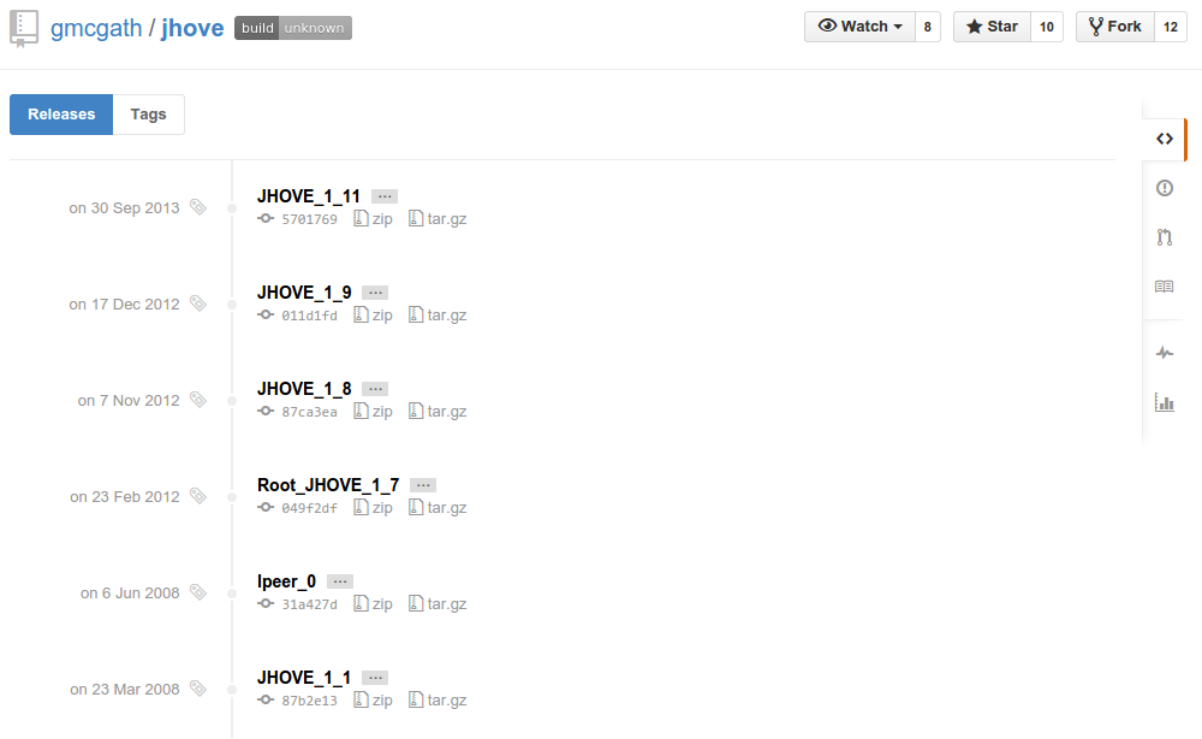
The only visible project documentation on GitHub is the README file which is an incomplete version of the SourceForge site tutorial. There are no links to the documentation hosted on SourceForge. The build instructions still reflect the ant build process despite that the ant build.xml file is no longer present.

Downloads

The GitHub downloads page lists a subset of JHOVE releases, these are offered as zip and gzipped tar archives. GitHub releases are marked by applying a Git tag and adding binary assets, release notes etc. An overview of GitHub releases and instructions to create them are online here.

GitHub doesn't provide a mechanism for tracking download statistics claiming that their attempted implementations were unreliable and adversely affected performance. Obviously downloads from the GitHub site might help to account for the decline in the SourceForge download statistics. There's no way to tell without access to the equivalent GitHub statistics.

The screenshot below is of the GitHub downloads page and shows the incomplete release list available.

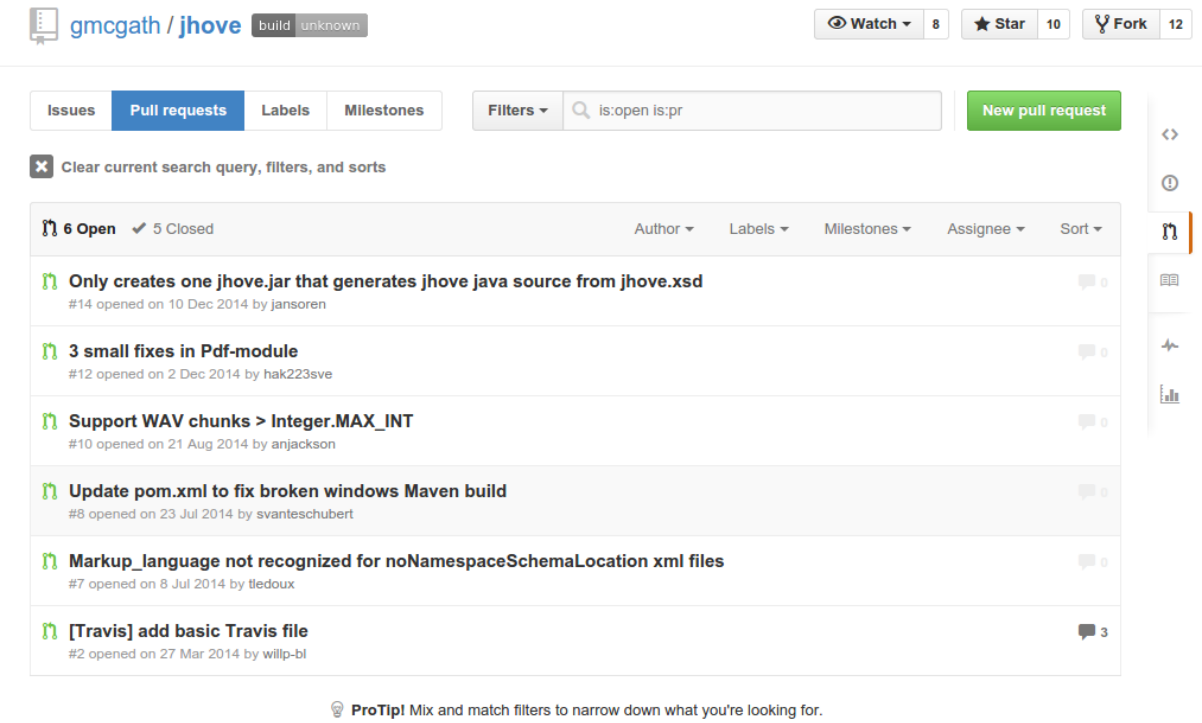


Issue Tracking

The JHOVE GitHub issue tracker is active here: https://github.com/gmcgath/jhove/issues. There are currently 2 open issues and 2 closed issues and no defined milestones. This leaves the open task of triaging these issues with the historic list on SourceForge. It's anticipated that a JHOVE developer / user community would take an active role in issue tracking and development roadmapping in future.

Pull Requests

There are currently six open pull requests for JHOVE on GitHub:



There are also 5 closed pull requests that have been merged since the 1.11 release.

Wiki

There is a single home page on the JHOVE GitHub wiki page that's effectively a status report created when JHOVE was transferred from SourceForge. The wiki lists the licence simply as GNU Lesser General Public License without stating a version.

JHOVE Release History

The table below lists all of the official releases of JHOVE with dates and corresponding SourceForge download names and Git tags shown on GitHub. Note that only a small number of releases have Git tags, there should be one for each official release.

Version		SourceForge Name	GitHub Tag
30/09/2013	1.11	JHOVE 1.11	JHOVE_1_11
22/09/2013	1.11beta	JHOVE 1.11b1	

10/06/2013	1.10	JHOVE 1.10	
04/03/2013	1.9	JHOVE 1.9	JHOVE_1_9
07/11/2012	1.8	JHOVE 1.8	JHOVE_1_8
05/08/2012	1.7	JHOVE 1.7	ROOT_JHOVE_1_7
04/01/2011	1.6	JHOVE 1.6	
23/12/2009	1.5	JHOVE 1.5	
31/07/2009	1.4	JHOVE 1.4	
04/06/2009	1.3	JHOVE 1.3	
11/02/2009	1.2	JHOVE 1.2	
28/04/2008	1.1	JHOVE 1.1	JHOVE_1_1

The last official JHOVE release was version 1.11 released 20/09/2013. The SourceForge package can be downloaded here, the tagged GitHub commit can be found here. This release was built using ant, subsequent development has altered the JHOVE build process to use Maven. There has yet to be a JHOVE Maven release, the project has been versioned as a 1.12 snapshot build.

External Resources

JHOVE XML Schema

The JHOVE XML schema file is available here: http://hul.harvard.edu/ois/xml/xsd/jhove/jhove.xsd, presumably hosted by Harvard. This logically matches the http://hul.harvard.edu/ois/xml/ns/jhove namespace for the schema.

Existing Debian Package

There is already an official debian package for JHOVE, the project tracking page can be found here: https://tracker.debian.org/pkg/jhove. The package hasn't been updated since December 2011, the packaged JHOVE version is 1.6 from January 2011. The tracker page acknowledges the existence of the 1.11 upstream version on JHOVE. The official maintainer is Jeff Breidenbach.

All of the JHOVE debian packages are distributed from the "non-free" repositories of their respective releases. The JHOVE 1.6 package is distributed from the current wheezy repositories and is also available from debian testing (sid), which will become the jessie release. The previous squeeze release non-free repository contain a JHOVE 1.4 package. These packages have also been distributed from the multiverse repository of recent Ubuntu releases:

- lucid (10.04 LTS) (JHOVE 1.4 squeeze package)
- precise (12.04 LTS) (JHOVE 1.6 wheezy package)
- trusty (14.04 LTS) (JHOVE 1.6 sid/jessie package)
- utopic (JHOVE 1.6 sid/jessie package)
- vivid (JHOVE 1.6 sid/jessie package)

Other Web Resources

There are a few other useful JHOVE resources available on the web that could / should be listed on an external resources page.

- JHOVE Usage Tips from Gary McGath: http://www.garymcgath.com/jhovenote.html
- Gary's File Formats Blog has multiple JHOVE posts: https://fileformats.wordpress.com/tag/jhove/
- A Digital Curation Centre JHOVE case study: https://www.era.lib.ed.ac.uk/handle/1842/3335
- Recent post by Yvonne On PDF validation with JHOVE: http://www.pdfa.org/2014/12/ensuring-long-term-access-pdf-validation-with-jhove/

Transfer of JHOVE Resources to OPF

This section presents the OPF's immediate plans for the transfer of the JHOVE project. It's anticipated that all of the tasks in this section will be completed for Friday March 27th. Once complete the OPF will be fully responsible for a curated set of the resources listed above. As a milestone it merely represents the establishment of a rationalised status quo. This is deliberate as it provides a stable foundation for the transition to community project.

Moving from SourceForge

The OPF's belief is that this is the time to end of life JHOVE's SourceForge presence. The division between the code base on GitHub with release packages, documentation and mailing list handled by SourceForge is trickier to administer and potentially confusing for newcomers. Whether this means clearly marking the project as dead and providing clear forwarding information, or closing the project is open to question.

The JHOVE Code Base

The SourceForge project is effectively dead as a working source repository with no activity since 2013. The full commit history has been transferred to GitHub already. The OPF will be using Gary McGath's GitHub project as the master source for the JHOVE code base.

The official OPF JHOVE repository will live at https://github.com/openpreserve/jhove.

Ideally this would be a transfer of Gary's repository, but the distributed nature of Git and GitHub means that this isn't necessary.

Branching Post 1.11 Release Development

The current JHOVE master branch contains post 1.11 development work, the history can be seen online here https://github.com/gmcgath/jhove/commits/master. Initially this work will be moved to a non-master development branch, or branches if it's easy to separate logically, leaving master at the 1.11 release of September 2013. This work will be assessed alongside the open pull requests as the code base is addressed post transfer.

Documentation

The current SourceForge documentation site is outdated and not optimally structured. The OPF proposes shutting down the site and providing clear forwarding to the new documentation resources.

The OPF will create two new documentation resources:

- http://openpreservation.org/technology/products/jhove, a hub page on the OPF's web site that presents:
  - an overview of the project and it's status;
  - where to obtain executable downloads;
  - simple, user-oriented "up and running" instructions; and
  - links to more detailed resources.
- http://jhove.openpreservation.org, a GitHub pages documentation site that carries more detailed information and how tos for users and developers.

The hub site will be administered by OPF and is meant as the point of first contact with the project. Once the transfer is complete and the release process for JHOVE finalised it's anticipated that the contents of the hubpage would be fairly static, with few changes

The GitHub pages site will be initially administered by OPF but any community member would be able to take this role. This is the replacement for the SourceForge site. The OPF has already transferred the existing SourceForge documentation to the new site that preserves the existing one.

We're currently reviewing the documentation to ensure it accurately reflects the current state of the project. Other minor issues such as layout, aesthetics and links to external resources will also be addressed.

Downloads

All of the binary downloads currently available on the SourceForge site have been made available through the OPF's JHOVE GitHub download page. The OPF has pinpointed the commits for all releases missing on GitHub and create an appropriate tag. The downloadable packages offered are identical to those offered on SourceForge. Releases have been augmented with GitHub release notes. The table below lists the JHOVE releases with their corresponding commit SHA1s and Git tag.

Version	GitHub Commit SHA1	Git Tag
1.11	57017690b114d4e4cf8b394b28bf52322084c2a2	v1.11
1.11beta	b703009fc0459717e11bbdc37df0e33454c94992	v1.11-beta
1.10	b5a030494e095ef4e69fd23c3e0f6f88d1c54f30	v1.10
1.9	011d1fd75260a2a84ae2e17eade27f805ec1fb4b	v1.9
1.8	d40dd00bb75381c15cd5df10983f620924dc545f	v1.8
1.7	ee7e47b467e392219c99aa6a06421e9d2dd14348	v1.7
1.6	7ff99d9cbe44f2f4bf61c753a3fbac16ea6f6a17	v1.6
1.5	1aeea61220e69e0a38049115f3e1ce82c9dc601e	v1.5
1.4	df7505687bf78bfba3668683e5cdba67a55c9668	v1.4
1.3	33c329f824375bd2a1819a9438fd272b744effe6	v1.3
1.2	c13dd52be0c7a83162cd6f37741f05d220576e97	v1.2
1.1	87b2e13cfc14d0b9bb6d7ebcf04fa4d193e30a05	v1.1

Initial plans for future JHOVE releases and downloads will be described in the OPF's technical evaluation of JHOVE, to be published separately. The intention for now is to ensure the longevity of current assets.

Issue Tracking

The OPF will triage the 25 open issues from SourceForge along with the 2 open GitHub issues. These will be tested while reviewing the current code base. Any issues that are current and relevant will be added to the OPF's GitHub issue tracker.

GitHub issues also supports development roadmapping through the use of milestones. JHOVE currently has no defined milestones.

Pull Requests

The 6 open pull requests will be evaluated as the OPF works towards a new JHOVE release that incorporates recent changes. For now they'll simply be transferred to the new repository.

Mailing List

The current mailing list hosted through SourceForge will be archived and discontinued. The OPF will set up a Mailman instance to host a new JHOVE mailing list. This also gives the OPF the capacity to establish and host other mailing lists.

Discussion Forum

There's no direct functional replacement for the SourceForge discussion forum on GitHub. A specific tag for OPF site blogs and comments seems an appropriate solution.

Wiki

The JHOVE project has historically managed without a Wiki meaning there's no transfer required. It might be that the move to genuine community project means that a shared documentation / discussion space for JHOVE would be useful. GitHub projects come with a free, built in Wiki or a space could be created on the OPF's Confluence Wiki.

External Resources

XML Schema

The official schema should remain at its current URL. The only question to be addressed is how the community would go about publishing any updates to the schema in future. One solution might be that future releases could use a community version of the schema, with a different namespace, hosted elsewhere.

Debian Packages

These are external resources and aren't affected by the transfer. Once a release process has been put in place for JHOVE (see later) the package maintainer should be contacted to see if they're interested in updating the existing package. An upstream maintainer of an official Debian package would be great to have in place once the project's house is in order.

External Resources

Again these aren't official project assets. The intention is to provide a curated list of these resources and link to them as appropriate in the new documentation.

A Community JHOVE Code Base

While the OPF takes over ownership of and responsibility for the JHOVE assets we'll also be putting in place some open source development infrastructure for the project. All of the tasks in this section will be complete by Friday 3rd April except the publication of Maven snapshot releases. That may take a little longer depending upon one or two administration issues.

Travis-CI

The OPF has set up two Travis continuous integration tasks for JHOVE. The first task builds the master branch, i.e. an ant build. This provides a reference point for regression testing new contributions for release. The second task builds the current development HEAD which is now an Apache Maven project. This build will form the basis of the next point release of JHOVE which will simply move the release and packaging to Maven without changing functionality.

This will provide an open source continuous integration service for JHOVE where all development branches are built and tested automatically whenever a developer pushes changes to gitHub. Any pull requests submitted to the official JHOVE project will also be built and tested automatically. Any developer forking the project to their own GitHub account would be able to use the projects Travis build file to set up a personal Travis build by flicking an online switch.

Unit & Integration Testing

The ant based build will not be amended to add any unit testing targets. Future JHOVE releases will be based on a Maven build so it makes sense to concentrate limited effort here rather than updating a legacy build. The OPFs initial efforts will be to establish automated JUnit testing as part of the Maven build. Specifically the OPF will be setting up the mechanism for running the tests as opposed to writing a full set of tests. The latter would require all of the effort available and more. The OPF will provide enough tests to provide standard mechanisms for adding and accessing test resources for the different HULs. The aim is to provide the community with something that can be extended following an established process.

OPF Jenkins

The OPF will also establish a nightly QA build of JHOVE on its Jenkins server. This isn't meant to repeat the role of the Travis tasks. Rather than been triggered by GitHub pushes this build

will be run on a nightly schedule and will:

- carry out static analysis of the JHOVE code base and publish the results via the OPF's Sonar server;
- build and publish a nightly Maven snapshot release of the current master branch; and
- publish project artefacts, e.g. JavaDoc and license analysis for the snapshot through a Maven site task.

Sonar Analysis

SonarQube is an open source platform for managing code quality. the OPF runs its own instance at http://sonar.opf-labs.org/. The purpose is to analyse the code base and report known problem patterns, for example complex/nested code paths. We're not suggesting that all of the issues raised by Sonar have to be addressed. Unit test coverage is likely to be limited for JHOVE as there's been no project policy that covers testing. The metrics generated provide hard data for making informed decisions regarding improvements and technical effort.

Publishing Project Artefacts

The OPF will publish a Maven site from the new JHOVE build, this will include Maven project details, up to date Javadoc and license details.

First Checkpoint

Once the above is in place the OPF will start publishing a nightly snapshot of the Maven project jars on the Maven Central repository. This is the OPF's first JHOVE release, but it's a developers release, rather than a user based one.

The end result is simply a stable platform to provide a base for moving JHOVE development forward. The OPF will be publishing a technical evaluation of JHOVE during April. This will also lay out options for possible future development / maintenance tasks.