

# Jhove 2: new modules for web archives

Clément Oury

Web archives preservation manager

Bibliothèque nationale de France

## What is Jhove (2) ?

- Jhove 1 is a format validation and characterization tool
- Evaluate the format of a file in order to ensure it will be possible to preserve it in the future
- Jhove 1 is widely used by heritage institutions
- Jhove 2: next generation tool
  - Core software developed by CDL
  - A module by module development for each new format

## BnF development project

- Jhove 2 modules necessary for ingest of our web archives in our digital repository ([SPAR](#))
  - ARC and GZIP modules
  - Unix “file” tool integration
  - Parallelism management to ensure scalability
- In-house development, outsourced to ATOS Origin
- Open source release

## How it works (basically)

Take an foo.arc.gz file and give it to Jhove2

- The arc.gz file is identified as a GZIP file by an external tool (Droid or File), then the GZIP format is validated and characterized by the Jhove2 GZIP module
- The file inside GZIP is identified as an ARC file by the external tool, then the ARC format is validated and characterized by the ARC Jhove2
- The format of each contained file is identified by the external tool, then this format is validated and characterized by Jhove2 (if relevant module available)

## Example: the ARC module

- Identification: currently through Droid
- What is a “valid” ARC file?
  - Version-block follows a fixed pattern
  - ARC-record structures comply with the version-block
- What features to extract?
  - Name, size, checksum of the ARC file; number of records
  - Context information from the version-block
  - Metadata from each record header
  - Information from protocol response (protocol version, result-code...)

## Outcomes and performances

- Jhove2 produces a highly configurable output with all possible information
- That can be mapped in any metadata scheme
- Performance depends
  - on the number of Jhove2 modules in use
  - on the amount of information you want
- Treatment of a standard BnF ARC file (180 Mo, 1000 records) = 20 secs.

## Current status

- Modules have been released by ATOS
- BnF validation process is underway
- Downloadable at:  
<https://bitbucket.org/lbihanic/jhove2-bnf>
- Documentation available upon request
  - Please try and test it!
  - Questions: clement.oury [at] bnf.fr