Performance Study of Digital Object Format Identification & Validation Tools

DLF

DIGITAL LIBRARY FEDERATION

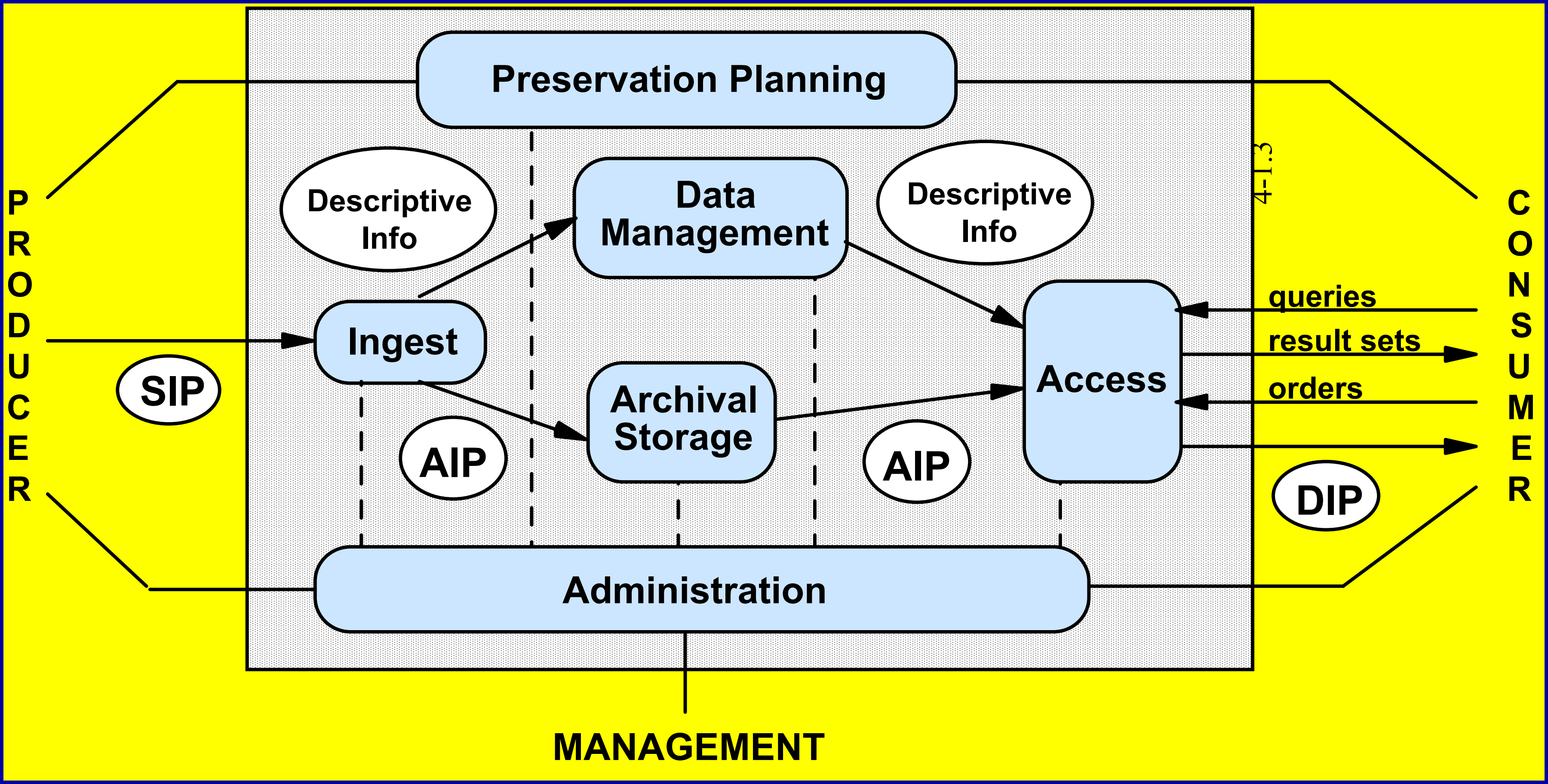
DLF Fall 2008 Forum
Nov 11-14, 2008

Quyen Nguyen
ERA Systems Engineering
National Archives & Records Administration

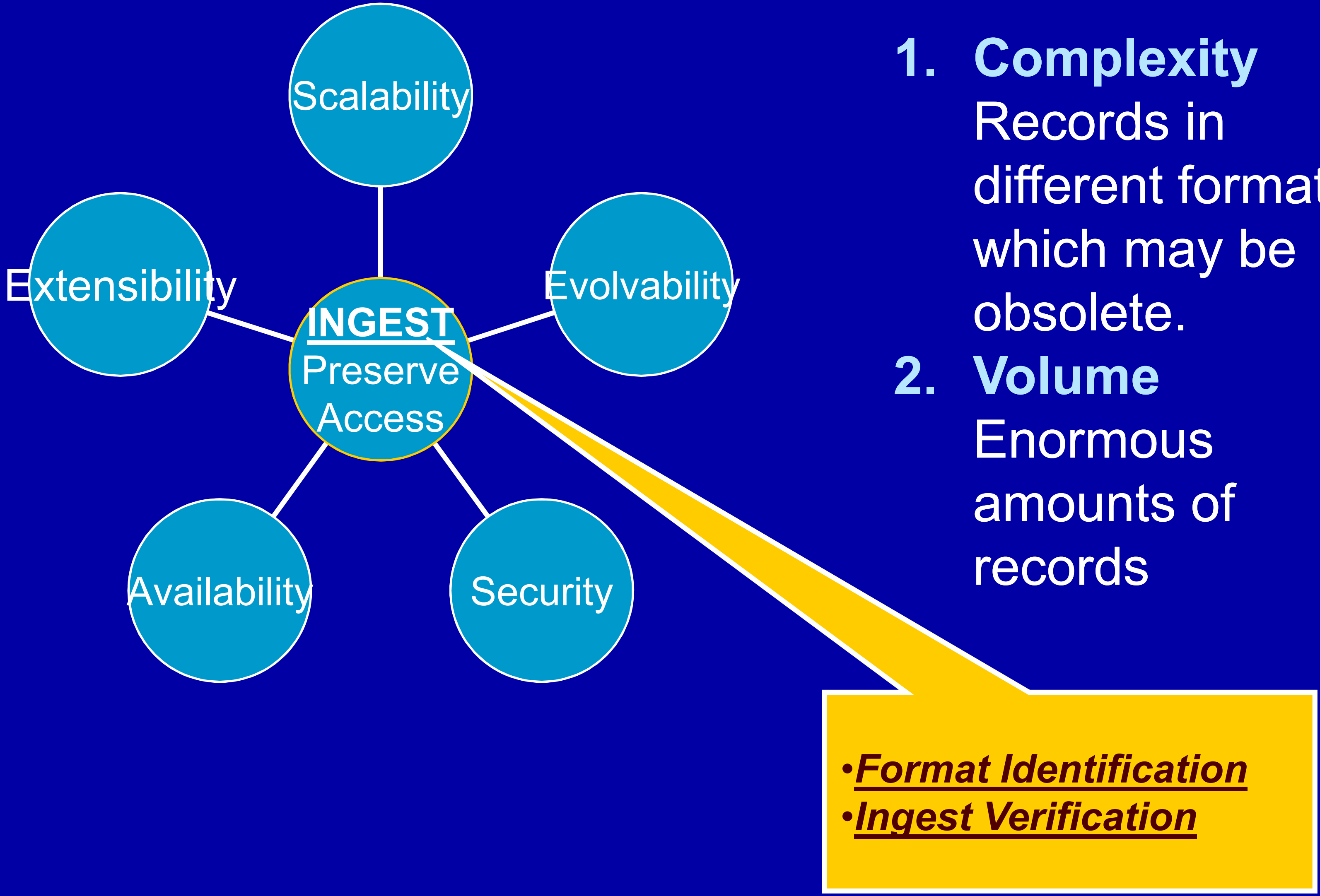
Agenda

- Background
- Format Identification Tools
- Experiments
- Analysis
- Related Work
- Summary

OAIS Model for ERA



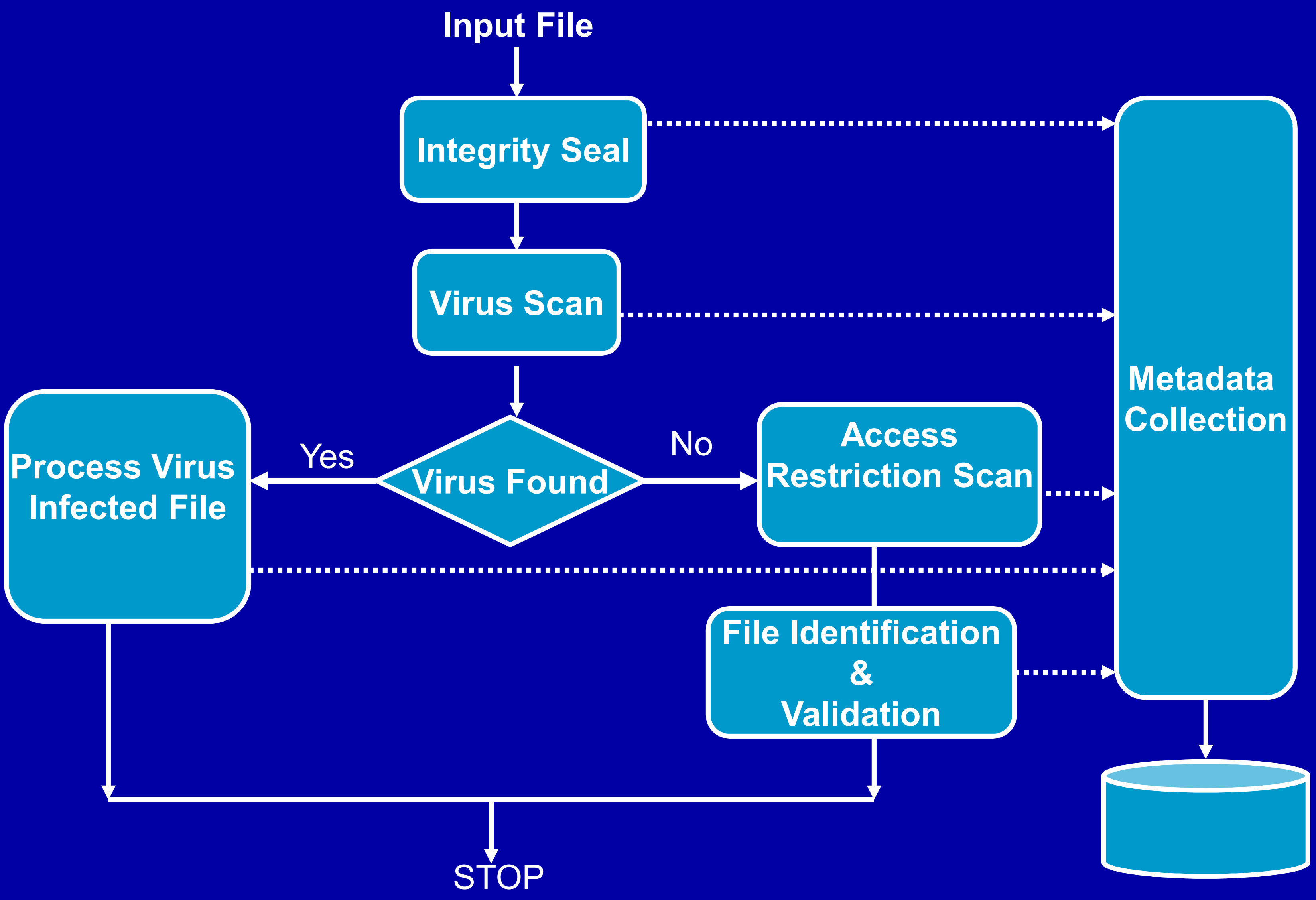
Challenges and Requirements



- ***Format Identification***
- ***Ingest Verification***

1. **Complexity**
   Records in different formats, which may be obsolete.
2. **Volume**
   Enormous amounts of records

Ingest Process Orchestration

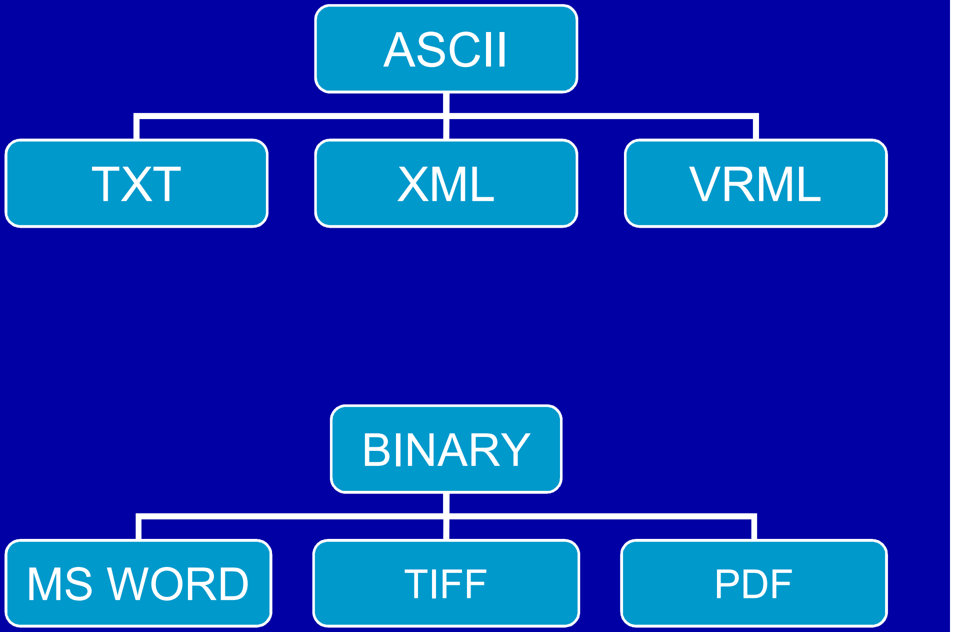


Agenda

- Background
- ➡ Format Identification Tools
- Experiments
- Analysis
- Related Work
- Summary

File Format

- Real issue: file extension unreliable to determine the format of a digital object
  - depends on end-user or application.
- *Format identification*. Microsoft Word 2003, Acrobat 8 PDF, etc.
- *Format validation*. Once a format *f* has been identified for a digital object X, does X really conform to format *f*. For example, an XML document may be well-formed or not.



Identification & Validation Tool

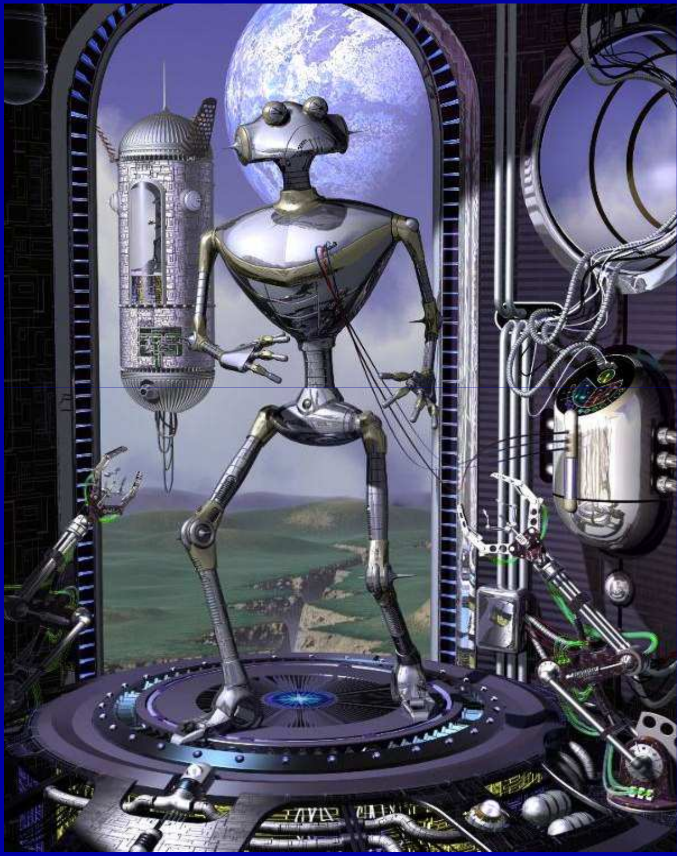
- Several institutions have developed such tools.
- A tool performs following task:
  - File → Input.
  - Find matched *signature*.
  - Output → Metadata:
    - File format: PDF, JPEG, Microsoft Word, etc.
    - Version number of application used to create digital object.
- Sounds simple yet difficult

JHOVE

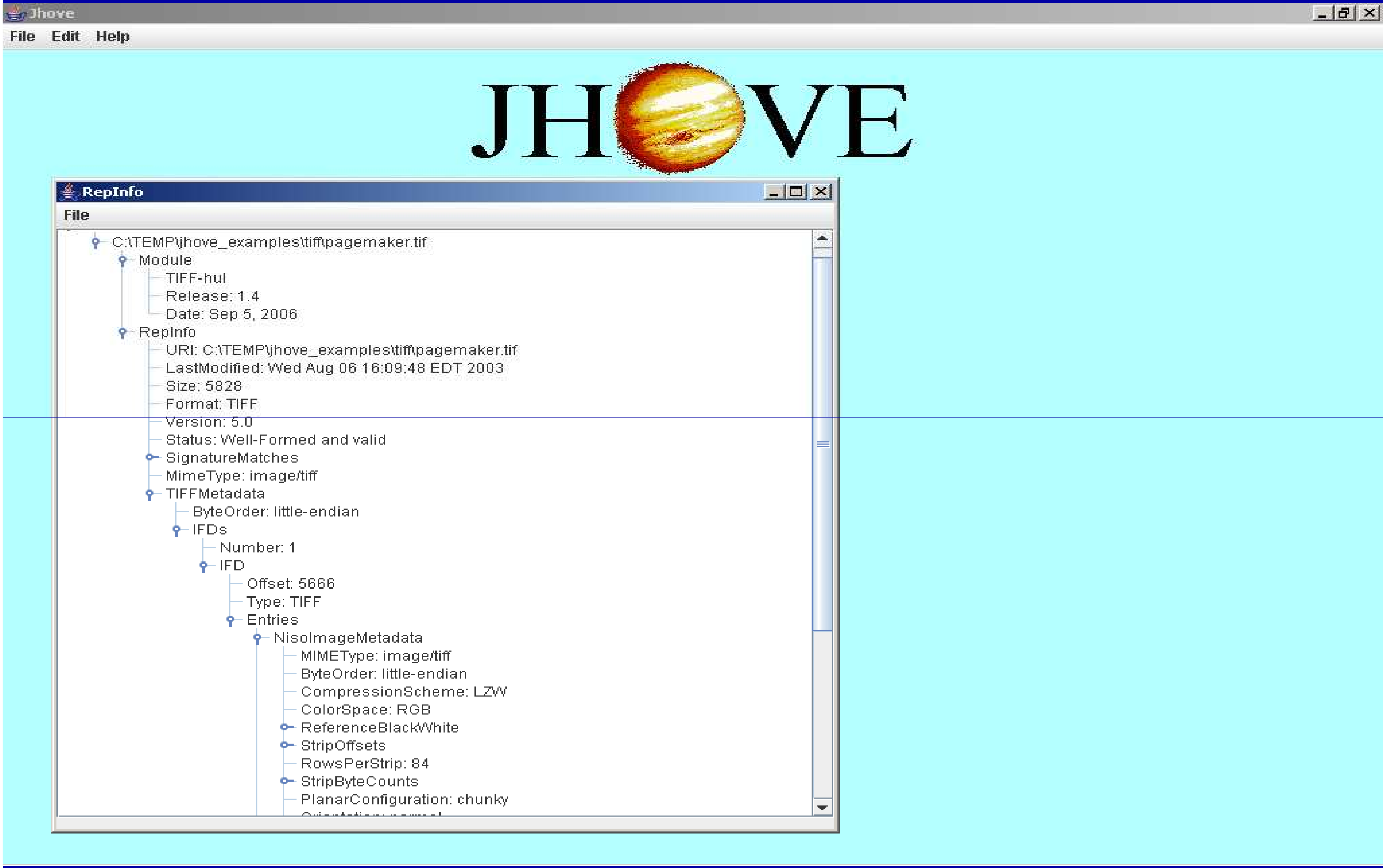
- **J**STOR/**H**arvard **O**bject **V**alidation **E**nvironment, developed by JStor (Journal STORage) and Harvard University Library.
- Set of modules called “handlers”, each of which is responsible for a file type
- Traverse set of “handlers” until one is found that can positively identify the type of input file.
- JHOVE can output rich metadata. Technical metadata such as MIX data elements for image files part of output.

DROID

- **D**igital **R**ecord **O**bject **Id**entification developed by United Kingdom National Archives.
- Based on PRONOM registry of file signatures specific to file types.
- At runtime, the content of the registry can be downloaded as an XML file, and cached in the DROID process.
- Traverse signature file containing cached content of PRONOM.
- DROID process will try to match one by one the signatures in the signature file against the one in the input file

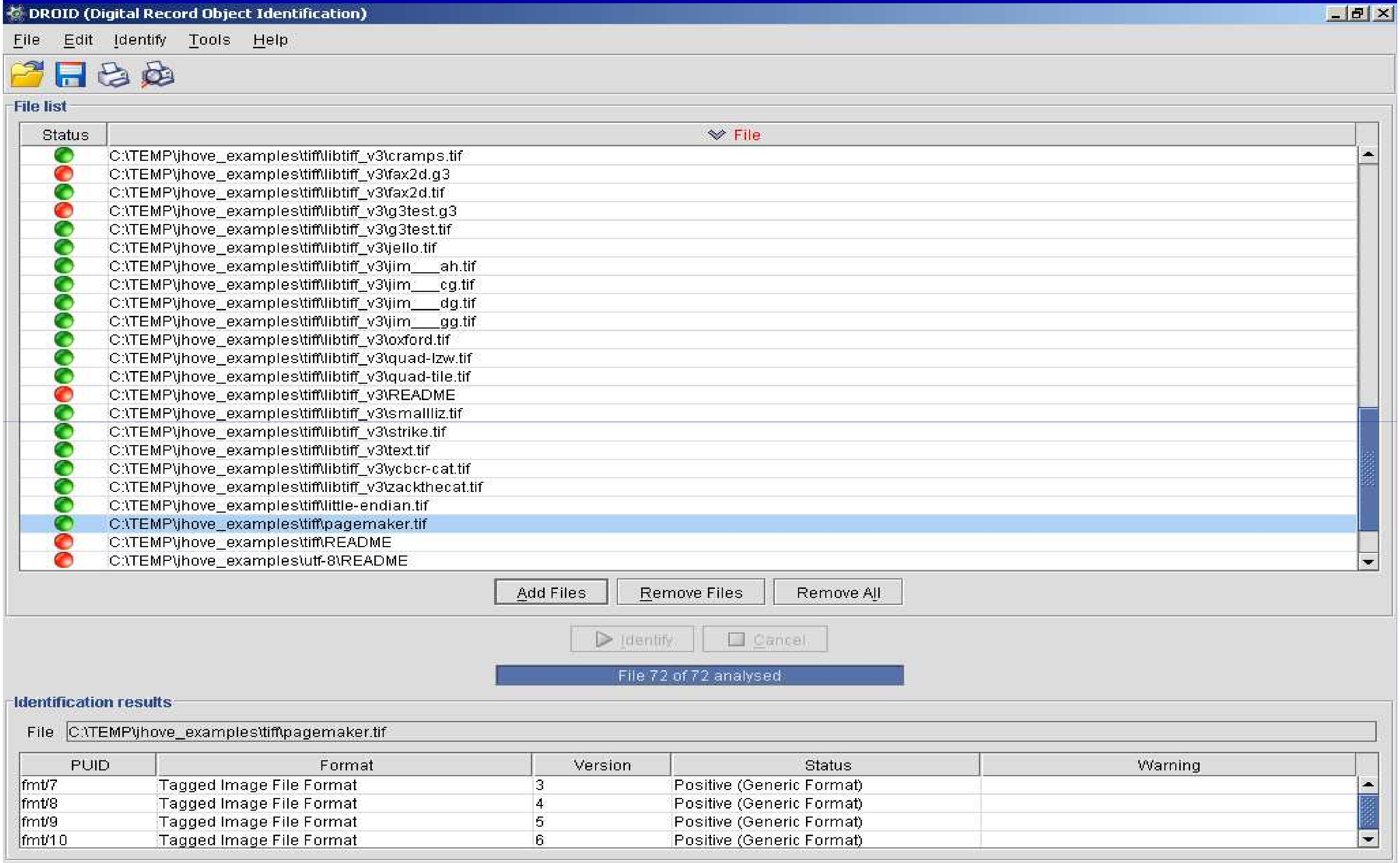


JHOVE Screenshot



```
C:\TEMP\jhove_examples\tiff\pagemaker.tif
    Module
        TIFF-hul
        Release: 1.4
        Date: Sep 5, 2006
    RepInfo
        URI: C:\TEMP\jhove_examples\tiff\pagemaker.tif
        LastModified: Wed Aug 06 16:09:48 EDT 2003
        Size: 5828
        Format: TIFF
        Version: 5.0
        Status: Well-Formed and valid
        SignatureMatches
        MimeType: image/tiff
        TIFFMetadata
            ByteOrder: little-endian
            IFDs
                Number: 1
                IFD
                    Offset: 5666
                    Type: TIFF
                    Entries
                        NisoImageMetadata
                            MIMEType: image/tiff
                            ByteOrder: little-endian
                            CompressionScheme: LZW
                            ColorSpace: RGB
                            ReferenceBlackWhite
                            StripOffsets
                            RowsPerStrip: 84
                            StripByteCounts
                            PlanarConfiguration: chunky
                            Orientation: normal
```

DROID Screenshot



Status	File
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\cramps.tif
🔴	C:\TEMP\jhove_examples\tiff\libtiff_v3\fax2d.g3
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\fax2d.tif
🔴	C:\TEMP\jhove_examples\tiff\libtiff_v3\g3test.g3
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\g3test.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\jello.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\jim___ah.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\jim___cg.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\jim___dg.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\jim___gg.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\oxford.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\quad-lzw.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\quad-tile.tif
🔴	C:\TEMP\jhove_examples\tiff\libtiff_v3\README
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\smallliz.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\strike.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\text.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\ycbcr-cat.tif
🟢	C:\TEMP\jhove_examples\tiff\libtiff_v3\zackthecat.tif
🟢	C:\TEMP\jhove_examples\tiff\little-endian.tif
🟢	C:\TEMP\jhove_examples\tiff\pagemaker.tif
🔴	C:\TEMP\jhove_examples\tiff\README
🔴	C:\TEMP\jhove_examples\utf-8\README

File 72 of 72 analysed

File C:\TEMP\jhove_examples\tiff\pagemaker.tif

PUID	Format	Version	Status	Warning
fmt/7	Tagged Image File Format	3	Positive (Generic Format)	
fmt/8	Tagged Image File Format	4	Positive (Generic Format)	
fmt/9	Tagged Image File Format	5	Positive (Generic Format)	
fmt/10	Tagged Image File Format	6	Positive (Generic Format)	

Agenda

- Background
- Format Identification Tools
- ⇨ Experiments
- Related Work
- Analysis
- Summary

Experimentation

- Environment
  - Intel ® CPU T2500 @ 2.0 GHz, 2.0 GHz, 2.0 GB of RAM.
  - Microsoft Window XP Professional Version 2002 Service Pak 2.
  - Runtime JVM comes with Sun JDK 1.6.0_01-b01.
  - java –Xms1024m –Xmx1024m
  - Jhove version 1.1 2006-02-13
  - DROID v1.1
- Inserted simple tracing code
  - `System.currentTimeMillis()`
  - `Runtime.totalMemory() - Runtime.freeMemory()`
- Metrics
  - Execution Time (ms): time-jhove, time-droid.
  - Heap Size (KB): heap-jhove, heap-droid.
- 50 measurements per collection or file.
- Statistical tools: Microsoft Excel and Stats4U.

Data Corpus

- Corpus C1: examples shipped with JHOVE.
  - 112 files whose size ranges from 1 KB to 22 MB
  - most of the files are less than 100 KB.
  - Files are grouped into subdirectories according to their document types: ASCII, GIF, HTML, JPEG, PDF, TIFF, WAV, and XML.
  - HTML subdirectory also contains GIF and JPEG images in the HTML pages.

Data Corpus (2)

- Corpus C2: 24 collections of documents used in NARA research lab.
- Typical documents coming to the public archives
  - Photos from National Park Service
  - Documents related to Katrina
  - Case files of U.S. District Courts
  - White House press releases,
  - Environmental maps from EPA
  - 1280 files whose sizes range from 1 KB to 136 MB.
  - Notably, document types are more varied.
  - In addition to the types found in C1 set, one can find audio, video clips files, geospatial files, statistical files, etc.

Statistical Analysis

- Perform T-Test using Microsoft Excel.

- *Execution Time*.
  - Null Hypothesis H0: time-droid = time-jhove.
  - Alternative Hypothesis H1: time-droid > time-jhove.

- *Heap Size*
  - Null Hypothesis H0: heap-droid = heap-jhove.
  - Alternative Hypothesis H1: heap -droid < heap -jhove.

- To conclude H0 with confidence, we want t-Stat be small, and P(T<=t) close to 1.
  - Watch for sign of t-Stat.

Experiment 1

- Corpus C1.

Execution Time:

- From T-test, 99% confidence level, time-droid is significantly greater than time-jhove
  - t Stat = 1487.29;
  - P(T<=t) one tail = 9.42E-182

Heap Size:

- From T-test, 99% confidence level, heap-droid is significantly less than heap-jhove
  - t Stat = -34.50825771
  - P(T<=t) one tail = 2.4239E-36

Experiment 2

- Corpus C2.

Execution Time:

- From T-test, 95% confidence level, time-jhove is significantly greater than time-droid
  - t Stat = -5.48;
  - P(T<=t) one tail = 2.51E-08

- DROID generated less heap memory than JHOVE

Data Type Impact

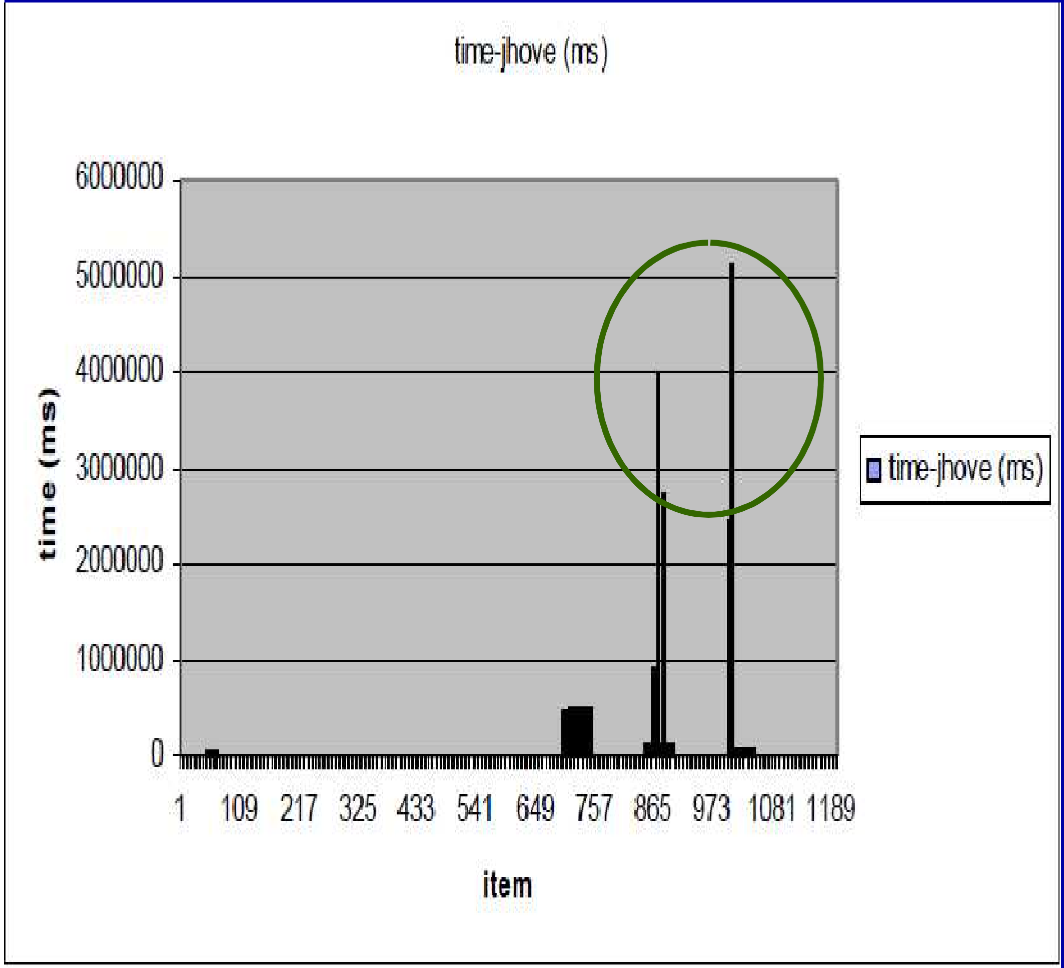


Figure 1. time-jhove vs. sample data points.

- Two collections -- around 865th and 1000th data points caused a dramatic increase in time-jhove
- Contain mostly VRML (Virtual Reality Modeling Language) files, which are essentially in ASCII text, but can be interpreted for display.

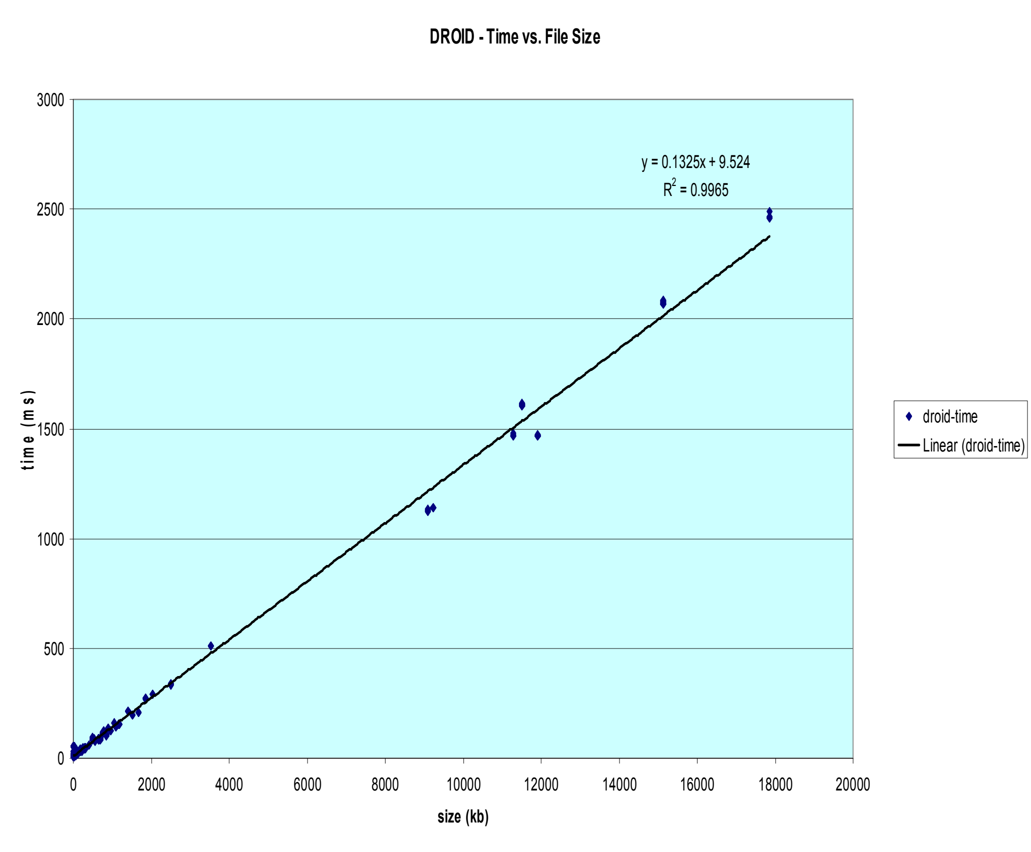
Experiment 3

- Tie.
- Corpus: C2b = C2 – {2 VRML collections}
- From T-test, 95% confidence level, time-droid is significantly greater than time-droid
  - t Stat = 0.057
  - P(T<=t) two tail = 0.95
- No difference on Heap size.

Experiment 4

- Corpus C2 re-arranged by types and sizes.
- Use Stat4U
- 3-way ANOVA with factors: A=Tool; B=Type; C=Size (2 levels only)
  - All 3 factors and interactions are significant with 95% confidence level.
  - Tool factor explains only 0.9 % of the variation.

Linear Regression: Size-Time



- Corpus C2.
- Only find linear regression for time-droid:

  time-droid = 0.13 * size + 9.52
- 100 TB → ~ 5 months.
- Information for sizing:
  - Computing resources
  - Parallelism

Agenda

- Background
- Format Identification Tools
- Experiments
- ⇨ Analysis
- Related Work
- Summary

Analysis

- Statistically, JHOVE and DROID perform equally well for format types that JHOVE can identify.
  - Qualitatively, JHOVE generated metadata is richer.
- For types that JHOVE cannot validate, the performance decreases drastically compared to DROID.
  - Easy case: if JHOVE finds that a record is binary, it just responds with a general identification, e.g. ByteStream.
  - But some ASCII cases such as VRML may throw it off.

Integrated Approach

- Two-phase approach for File Identification and Validation:
  - Pass a file through DROID to quickly identify its type.
  - If the type is found to be on the known list of JHOVE, then pass through JHOVE to extract technical metadata.
- These extracted technical metadata useful for automatic verification purposes.
- Examples include image resolution, format version numbers, creation dates, font information, etc.

Agenda

- Background
- Format Identification Tools
- Experiments
- Analysis
- ⇨ Related Work
- Summary

Related Work

- GDFR: Global Digital Format Registry
  - Distributed and replicated registry of format information
  - Allow the registration and discovery of digital formats for the long term
  - Collaboration of Harvard University Library and OCLC
  - http://www.gdfr.info

- FOCUS: Format Curation Service at the University of Maryland
  - Main component is Format Identifier (Fider)
  - Registry Global Digital Format Registry (GDFR) implemented using LDAP
  - https://wiki.umiacs.umd.edu/adapt/index.php/Focus:Main

Related Work (2)

- PERPOS: Presidential Electronic Records Pilot System at Georgia Technology University
  - software tools to support the OAIS functionalities
  - http://perpos.gtri.gatech.edu
- Metadata Extract Tool from National Library of New Zealand
  - http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool
- AIHT: Automated Preservation Assessment of Heterogeneous Digital Collections bys Stanford University
- The University of London Computer Centre issued a report to compare DROID, JHOVE, and AIHT:
  - *Assessment of File Format Testing Tool*. http://www.ulcc.ac.uk/uploads/media/DAAT_file_format_tools_report.pdf.
  - Very good qualitative and functional analysis
- Other Projects?

Summary

- File Identification and Validation important step in Ingest Process.
- Performance study of Jhove and DROID.
- Optimal approach leveraging both tools.
- Monitor future progress of other tools.
- Looking forward to Jhove 2.

Thank You

http://www.archives.gov/era

For any comments or questions, please mailto: quyen.nguyen@nara.gov