Evaluation of format identification tools

Johan van der Knijff

Koninklijke Bibliotheek – National Library of the Netherlands

johan.vanderknijff@kb.nl

The Future of File Format Identification
DPC / The National Archives, Kew, London, 28.11.2011

Background & context

About SCAPE

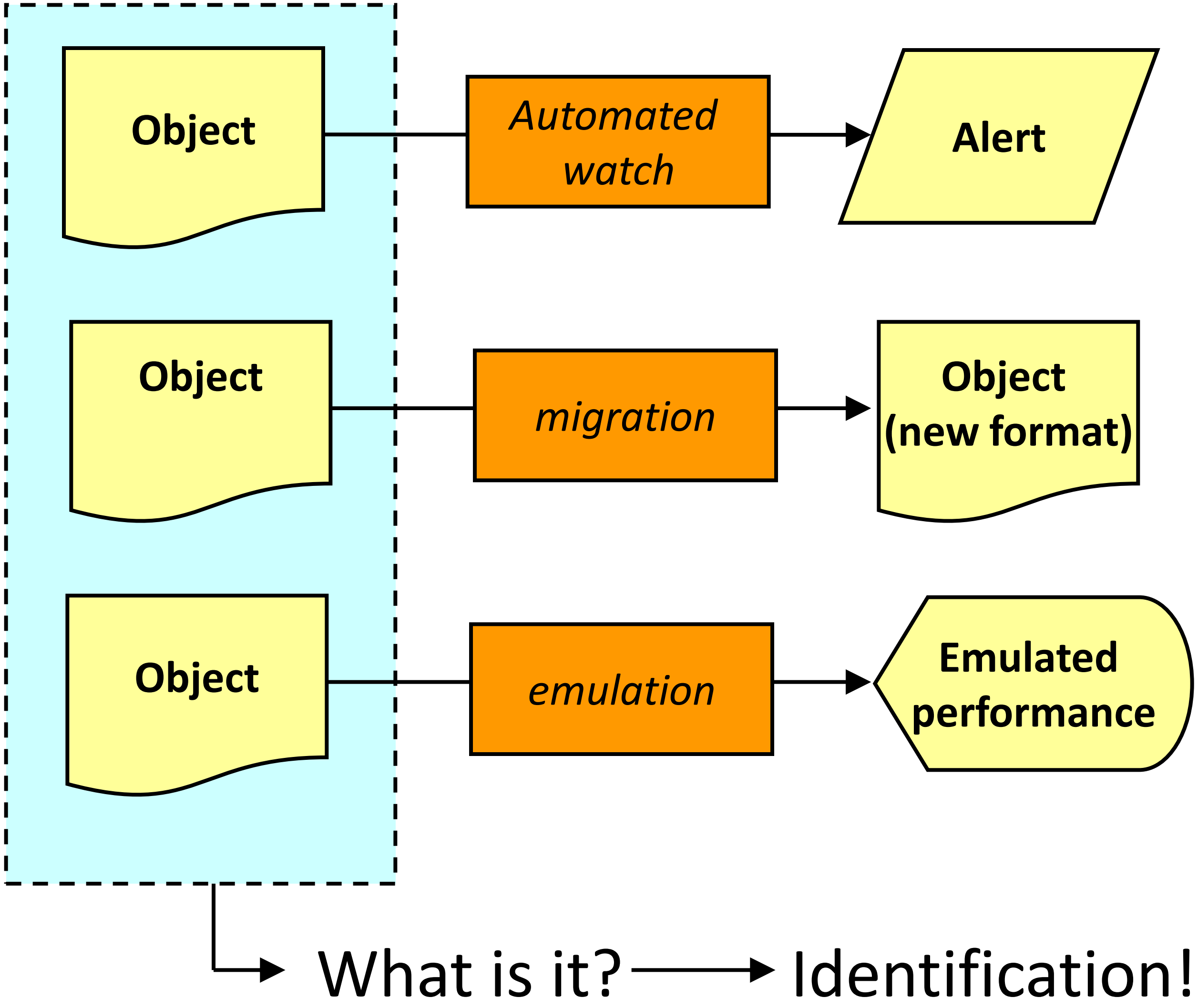
SCAlable **P**reservation **E**nvironments

EU funded FP7 project; 16 partners

Scalable services for preservation and preservation planning

Semi-automated workflows for large-scale, heterogeneous collections of complex digital objects

Importance of identification



Object

Automated watch

Alert

Object

migration

Object (new format)

Object

emulation

Emulated performance

What is it? → Identification!

Evaluation of identification tools

Which tools suitable for SCAPE architecture?

Specific strengths/weaknesses

Decide on needed enhancements and modifications

Hopefully provide some useful input to developers as well!

Tools

DROID 6.0	The National Archives
FIDO 0.93/0.95	Open Planets Foundation
Unix File Utility 5.0.3	unix
FITS 0.5 (uses DROID 3.0)	VE RI TAS HARVARD
JHOVE2 (uses DROID 4.0)	JHOVE2

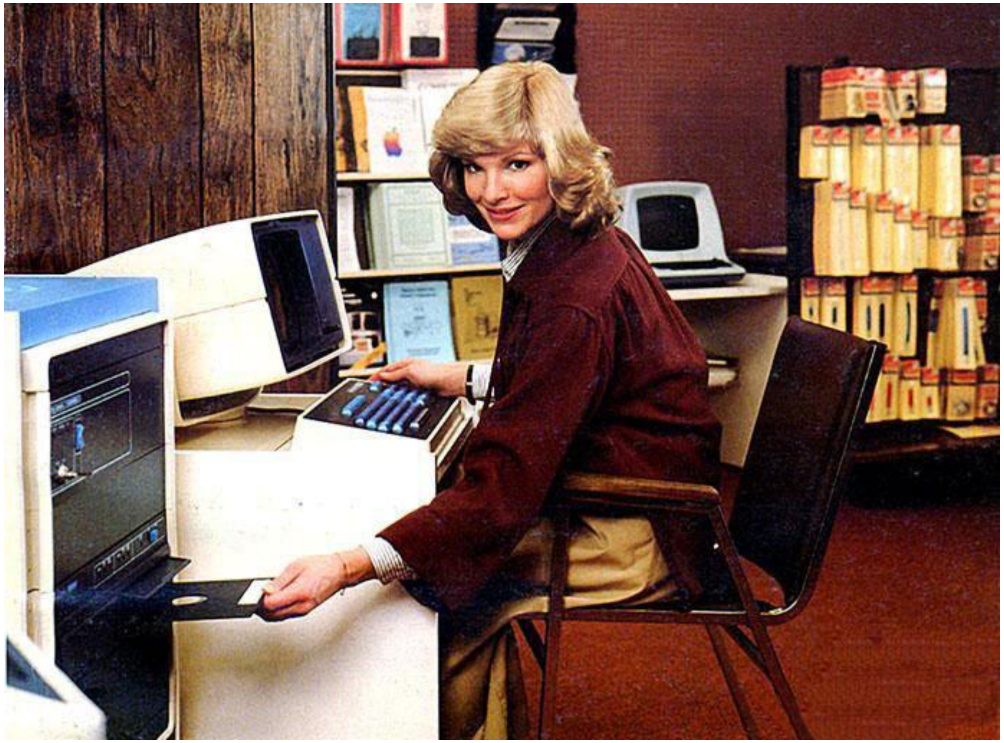
Evaluation framework

Total of 22 criteria, broadly covering:

- Usability in automated workflow (interface, dependencies)
- Fit to requirements archival setting: format coverage, extendibility, accuracy
- Output: format, identifiers, granularity
- User documentation
- Performance, stability and error handling

Key principles

Hands-on testing using real data is essential!

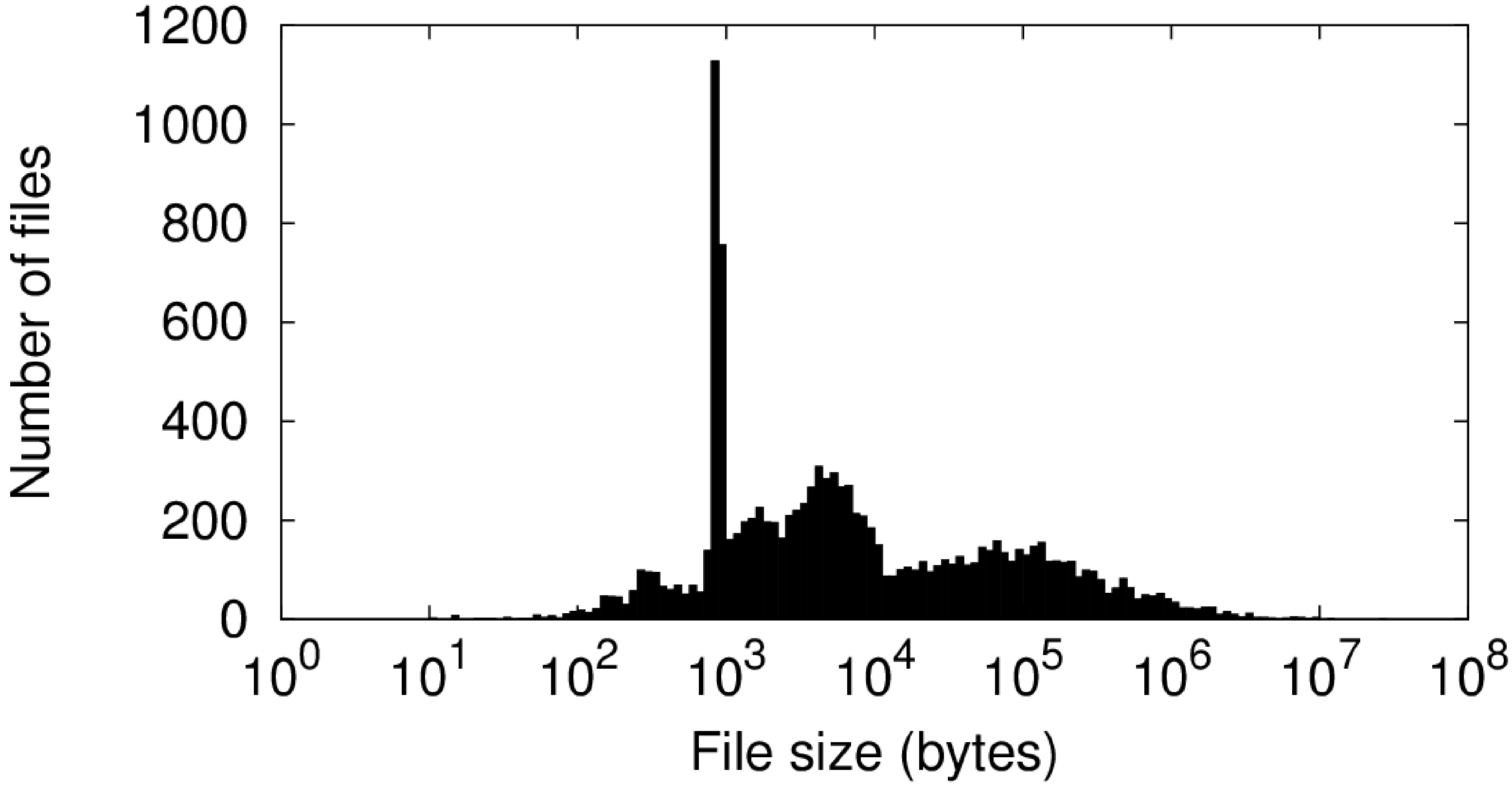


Key principles

Inform tool developers on results, and give them opportunity to provide feedback

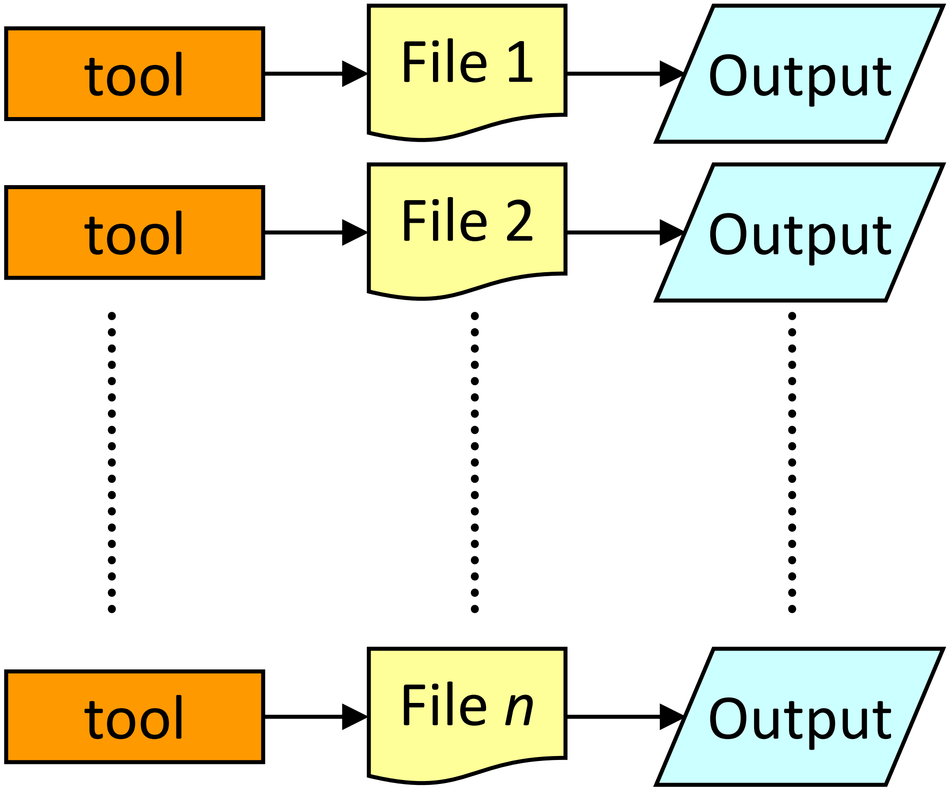
Performance tests

Test data: KB Scientific Journals set

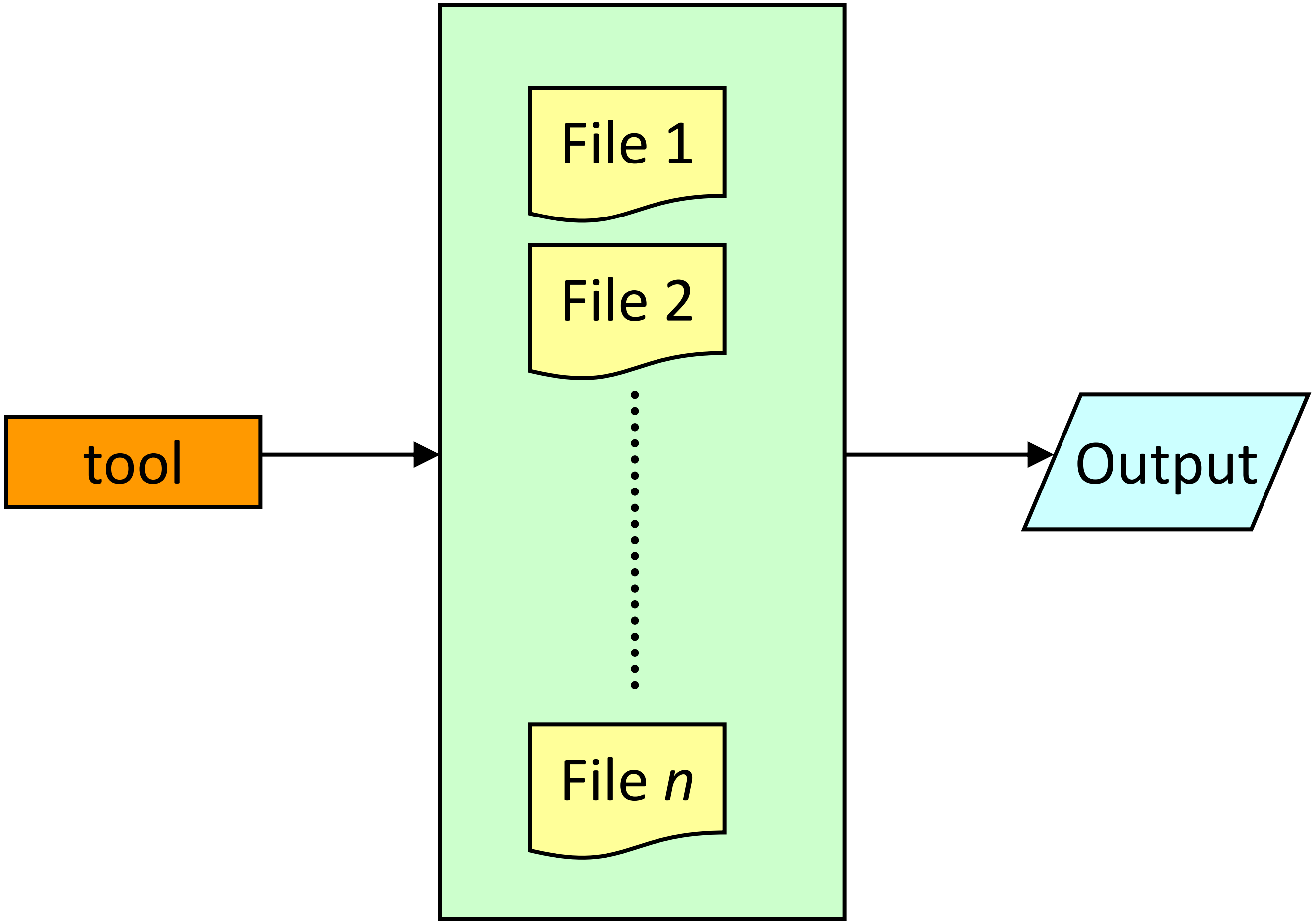


N	size(min)	size(median)	size(max)	Total size
11,892	11	4,737	25,495,289	1.15 GB

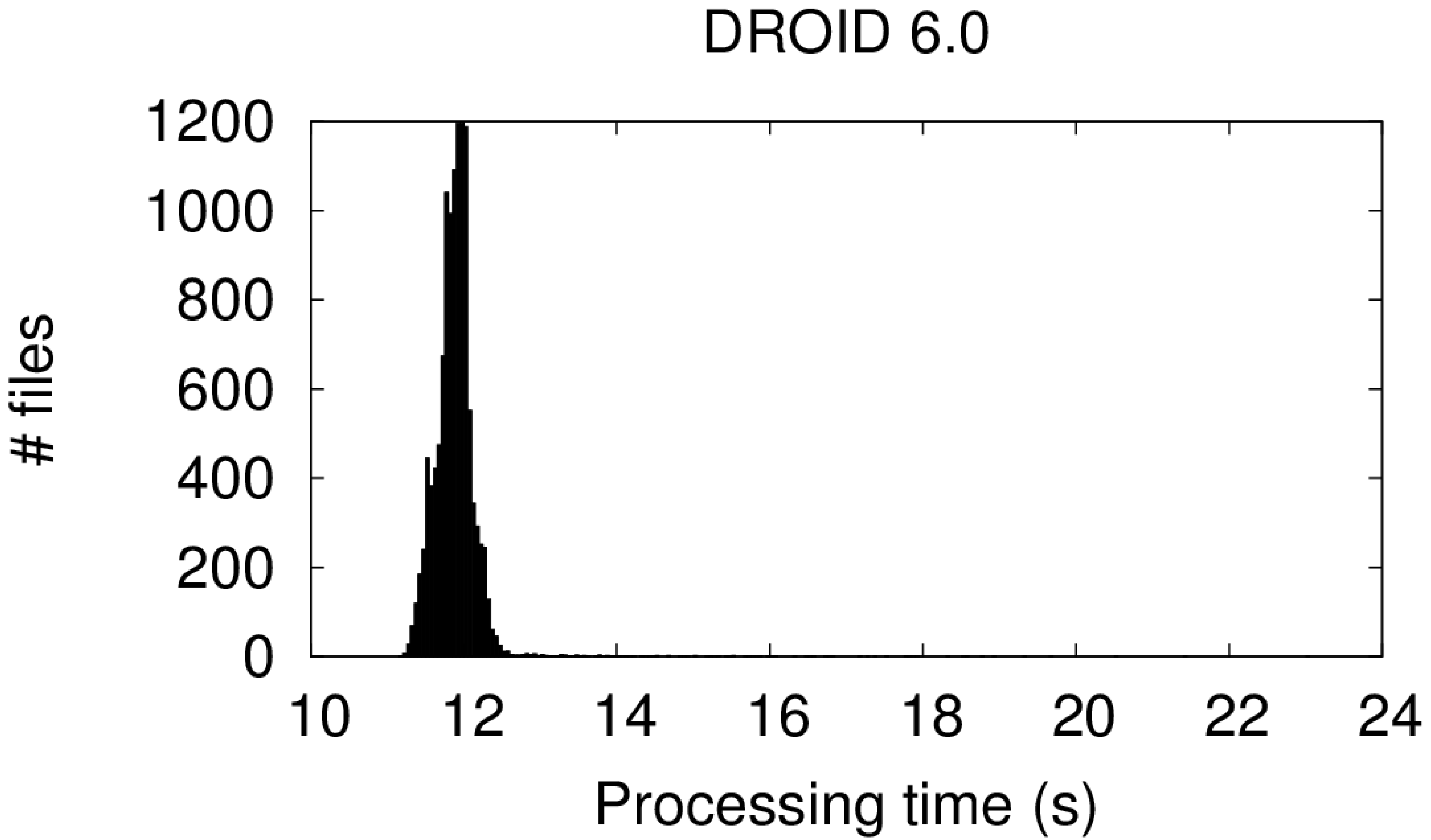
One file per tool invocation



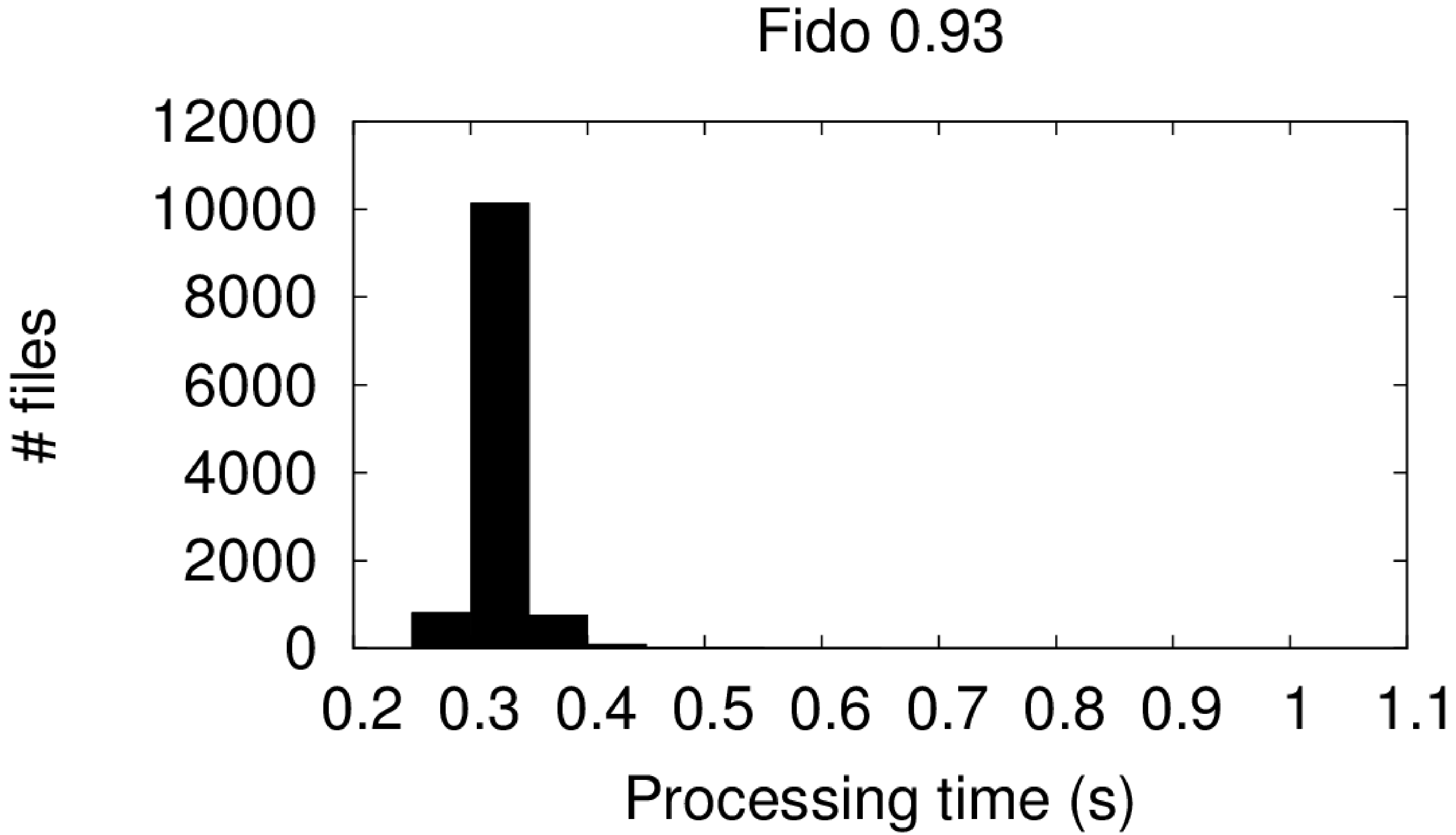
Many files per tool invocation



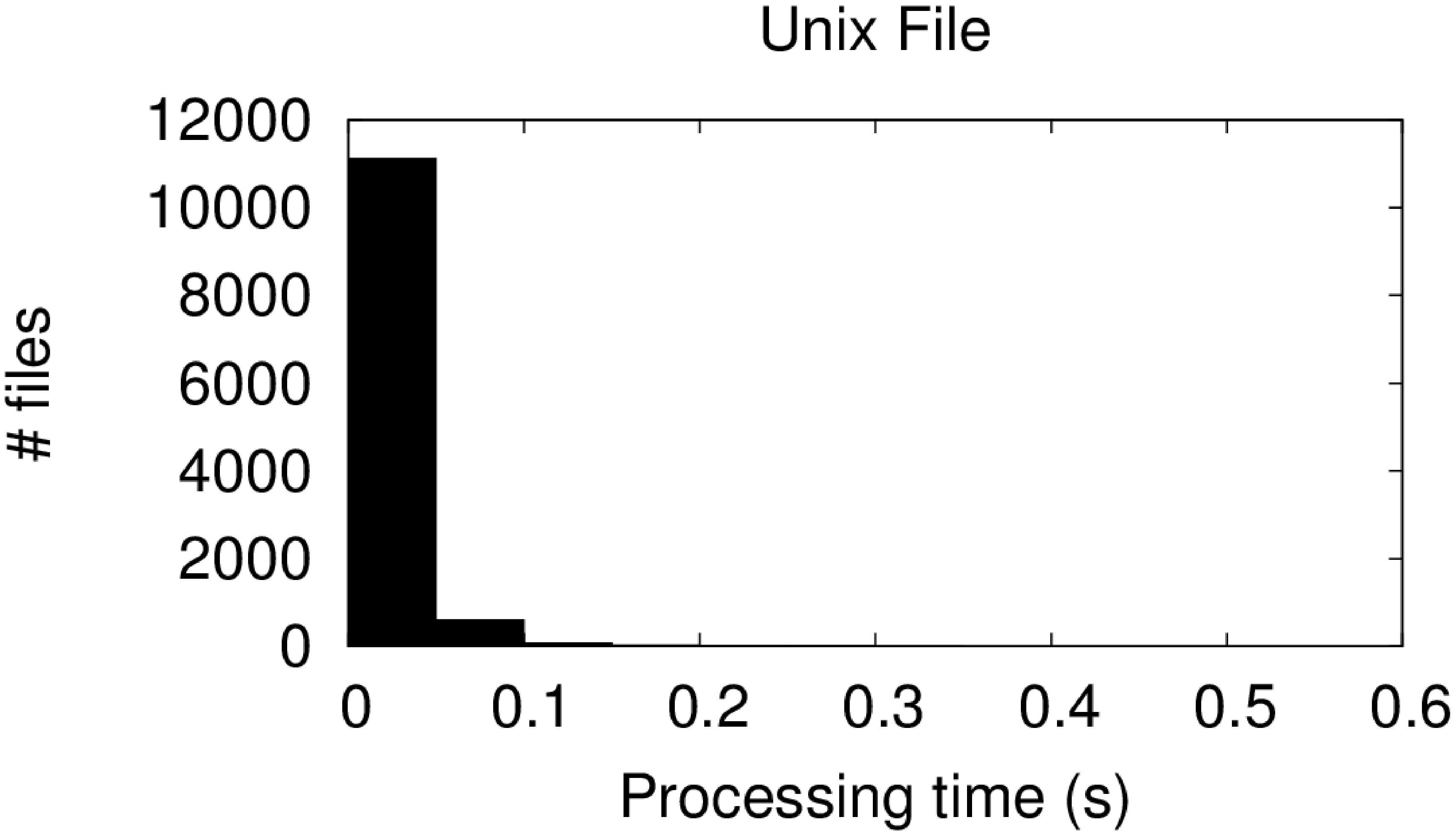
One-file per invocation



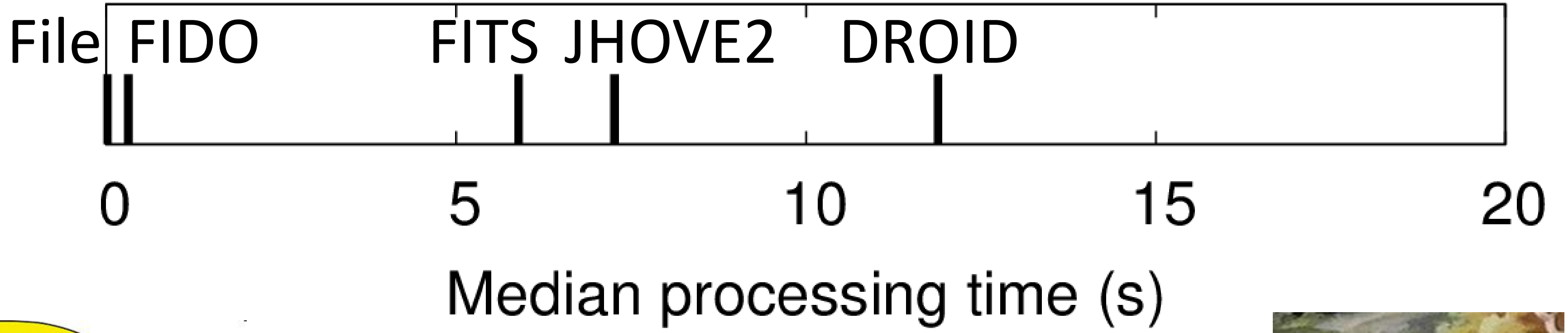
One-file per invocation



One-file per invocation

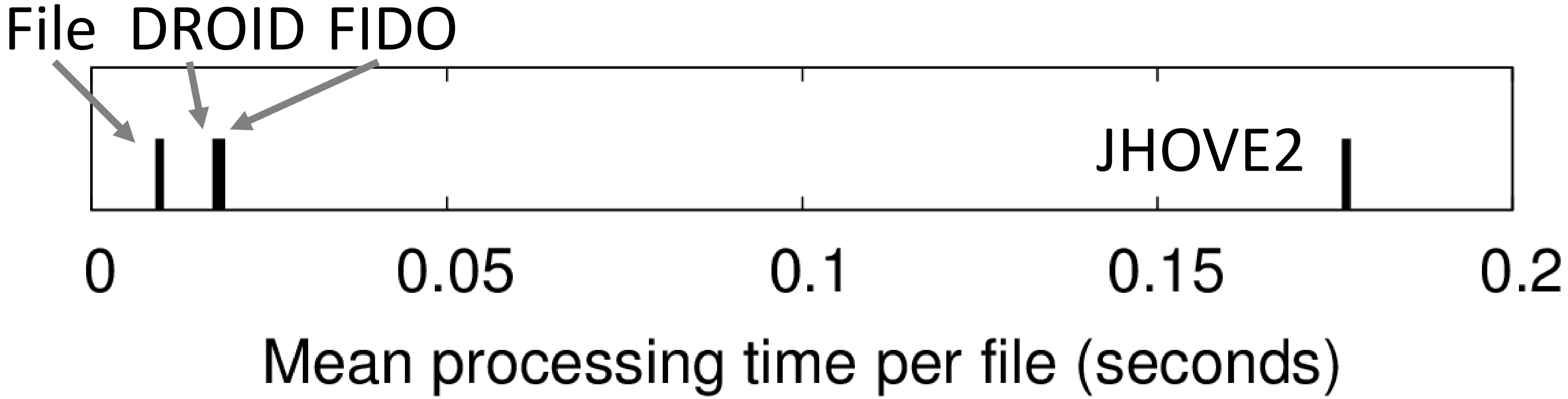


Comparison: one-file per invocation



SCAPE

Comparison: many-files per invocation



Performance: main conclusions

All tested Java-based tools slow for one-file-per-invocation use case

Performance much better for many-files-per-invocation use case

Slow initialisation seems to be main culprit

- Actual processing time per file: milliseconds
- Tool initialisation time: several seconds!

So is this really a problem?

Depends on required throughput

Depends on workflow interface (command line or Java API)

Depends on organisation of workflow

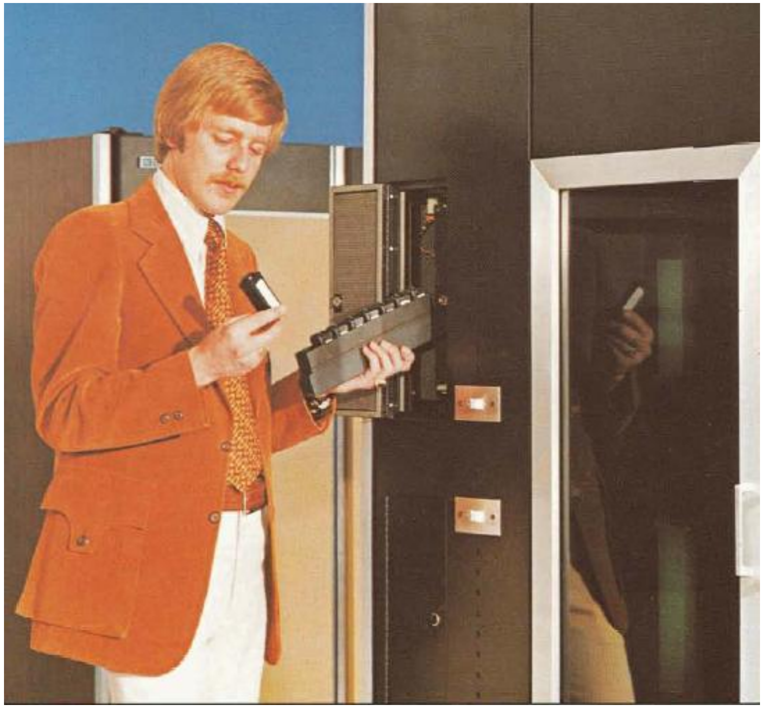
Depends on purpose (e.g. pre-ingest vs profiling of large file collections)

Apples vs oranges

FITS, *JHOVE2*: wrappers; also feature extraction and validation

DROID 6, *FIDO*: recurse into ZIP files ; *File* doesn't!

Miscellaneous observations



Other observations

Signature-based identification doesn’t work too well for text-based formats (including XML)

File outperforms other tools on format coverage and performance; management of signatures (‘magic’ file) awkward

DROID 6 output handling clumsy in automated workflows (separate *DROID* invocation needed for exporting profile information!)

Response to this work so far

FIDO: version 0.9.6 released in October; fixes most reported issues

FITS: version 0.6 released in October; various enhancements based on outcome of evaluation

DROID, JHOVE2: both provided feedback and will consider test results for upcoming releases

Possible next steps

Improve evaluation of accuracy

Keep up with tool updates; keep this work up-to-date

Publish all used scripts and detailed description of analysis methods so others can contribute more easily

Use publicly available test corpus (e.g. *Govdocs1*)

Link to full report on OPF blog:

www.openplanetsfoundation.org/blogs/2011-09-21-evaluation-identification-tools-first-results-scape

Open Planets Foundation

More about SCAPE:

http://www.scape-project.eu

twitter #SCAPEProject