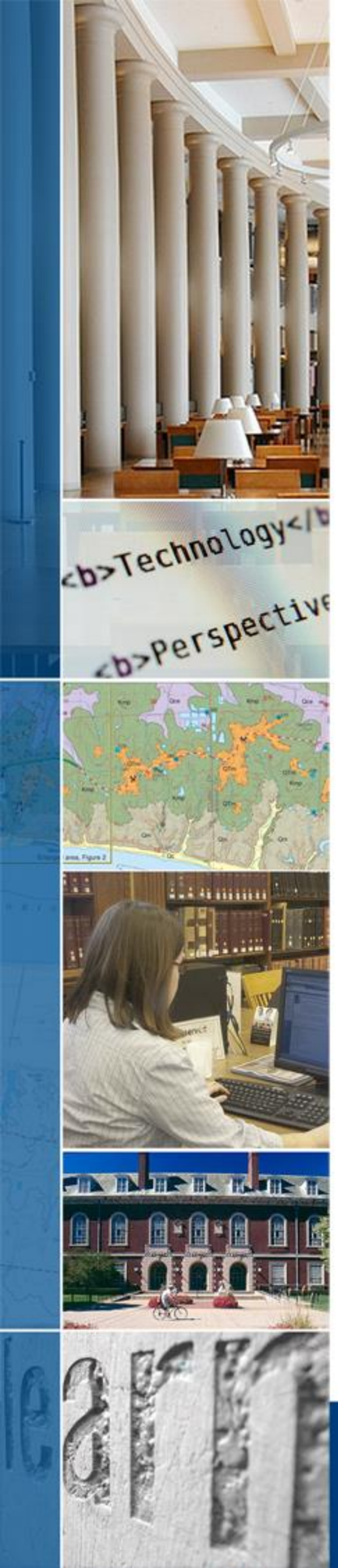


Advancing File Format Policymaking for Digital Preservation at the University of Illinois at Urbana-Champaign

Kyle Rimkus and Scott Witmer

Preservation Services, University Library, University of Illinois at Urbana- Champaign

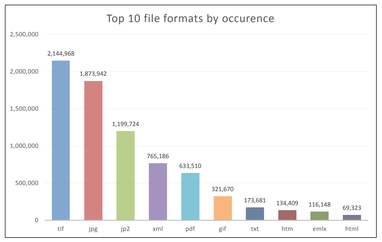
Abstract

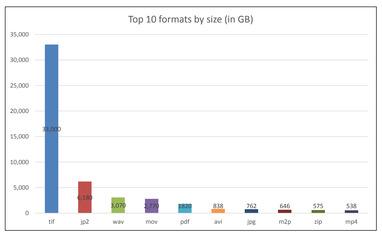
This research seeks to advance digital preservation theory and practice by presenting an evidence-based model for file format policy management in digital repositories. Its authors demonstrate a practical method for expressing and testing file format policies in association with software and operating system profiles which can be used to produce a statistically valid overview of preservation risk to digital collections.

What is Medusa?

- Medusa is the UIUC Library's digital preservation repository.
- Medusa is home to over 8,000,000 master files from digital collections of enduring value.
- At present, Medusa's collecting focus is on digitized and "born digital" books, manuscripts, photographs, audiovisual materials, scholarly publications, and research data from the library's special collections, general collections, and institutional repository.

Files in Medusa





What is a File Format?

A file format provides structure for the encoding of digital files for computer storage. Different file formats represent different ways of organizing data for particular forms of media. Digital preservationists are concerned with identifying file formats that are likely to remain accessible over time.

What is a File Format Profile?

File format accessibility depends on the ability of rendering applications and operating system environments to read and display the files. For each file format in Medusa, a format profile captures information about the platforms necessary to open files in that format. Profiles also contain information about file extensions and types of media content.

File Format Profile: JPEG

Status: active
Software: Adobe Photoshop
Software Version: 14
Os Environment: Windows
Os Version: 7
Notes:

} rendering software
} operating system

Content types:	File Extensions:
image/jpeg	jp
	jpe
	jpeg
	jpg

media / MIME type

Testing Process

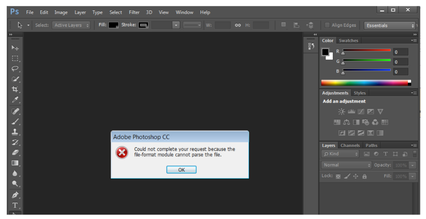
- Generate a random file in Medusa.
- Run FITS, a file format identification and validation tool, to extract technical metadata for future analysis.
- Download the file and open it using software identified in the file format profile.
- Post results of test to file information. If file fails to open, note issues or error messages and select reason for test failure.

In future rounds of testing, we will attempt to open failed files using alternate applications.

Problems

JPEG 2000 (.jp2)

- A majority of test failures so far are JPEG 2000 files that were created during the initial adoption period by an outside vendor.
- These files are recognized as valid, but the file header contains information that cannot be parsed by current versions of image editing software.
- Further investigation will determine which batches of JPEG 2000 files need to be diagnosed for this issue and potentially converted to another format.



File information

Storage level: bit-level
MD5: 8fcb20802edc811bd1c2241a3b5354fe
UUID: 2aacede0-5cd2-0132-3334-0050569601ca-5
FITS:
Belongs to: 2007-2014 Master Files ("Archive" folder)
File size: 92 KB
Mimetype: image/jp2


ARGUMENTS OF BOOKS I. TO XII. OF THE ILIAD.

BOOK I.

BOOK II.

File Format Test

Tester Email: sdwitme2@illinois.edu
Date: 2015-10-23
File Format Profile: JPEG2000
Status: Fail
Notes: Renders in Medusa Kakadu, not in Photoshop.
File Format Test Failure Reasons
- Software's file format module cannot parse the file

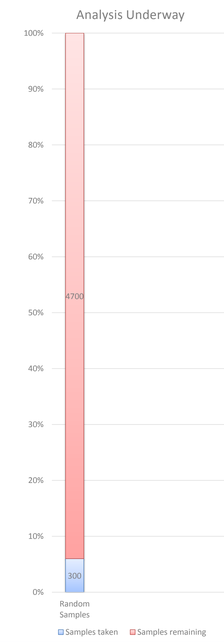
Incomplete or Unknown files

- Certain file formats require other files to be completely rendered.
- Files that are missing extensions or files with unrecognizable or conflicting media types lack the information necessary to open the files.
- Some files are temporary process data used by an application during the creation of another file. Files containing data elements that supplement another file become unidentifiable when taken out of context.

Status

All results are preliminary!

- Research is currently in the data collection phase.
- Testing began in October 2015 and is scheduled to end January 2016.
- To date we have tested ~300 files out of a target testbed of 5000 samples needed for results to be within a 2% margin of error.



Goals

We will triage files not only by format, but by provenance, in order to identify trends in different populations of data:

- All files in Medusa by format
- All externally deposited IDEALS files
- All internally digitized and deposited IDEALS files
- All University Archives "born digital" files

This research will be summarized in a paper when completed, and targeted for presentation at iPres 2016.

Acknowledgments

The authors would like to thank the University of Illinois at Urbana-Champaign Campus Research Board for generously supporting this project.

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA CHAMPAIGN