PURDUE UNIVERSITY Libraries Access. Knowledge. Success.

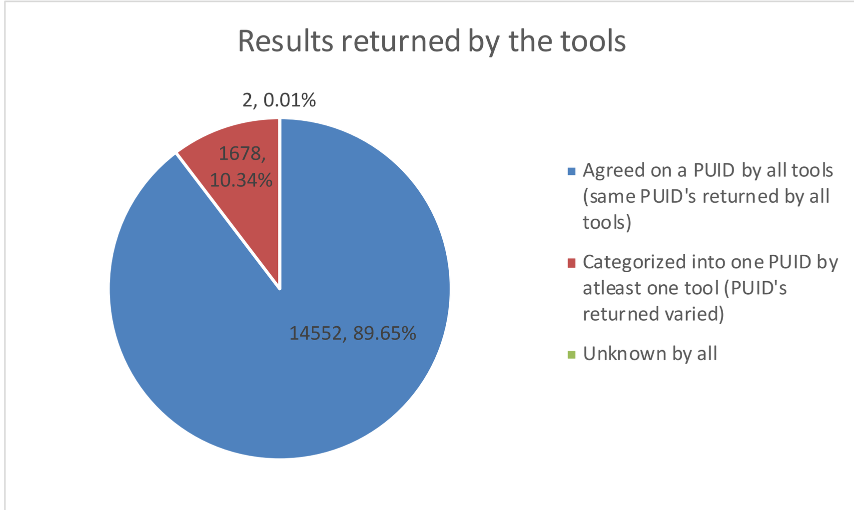
It is a common practice to use profiling tools to automate the identification of the formats of files that are submitted to a repository, to populate technical metadata describing the files, and to assign and execute different preservation actions that are based on the formats of the files. Incorrectly identifying or failing to identify a particular format represents a critical fault in most, if not all, digital preservation strategies. When problems with a widely adopted tool were encountered by the staff of an institutional data repository in the United States, an experiment was conducted to compare the results of a set of profiling tools [FIDO, Siegfried, DROID, and DROID in profile mode] that were scripted to ascertain the PRONOM Unique Identifiers (PUID) for 16,232 files in the repository.

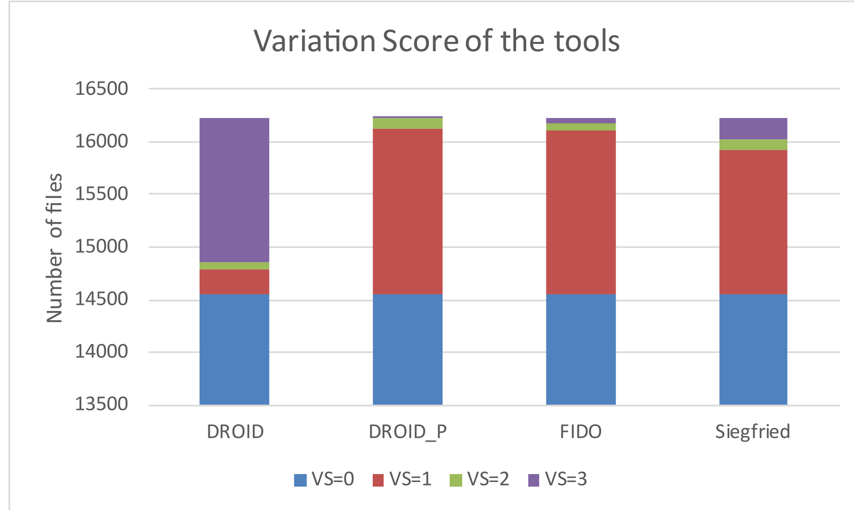
FIDO (Format Identification for Digital Objects) is a command-line tool developed by the Open Preservation Foundation.

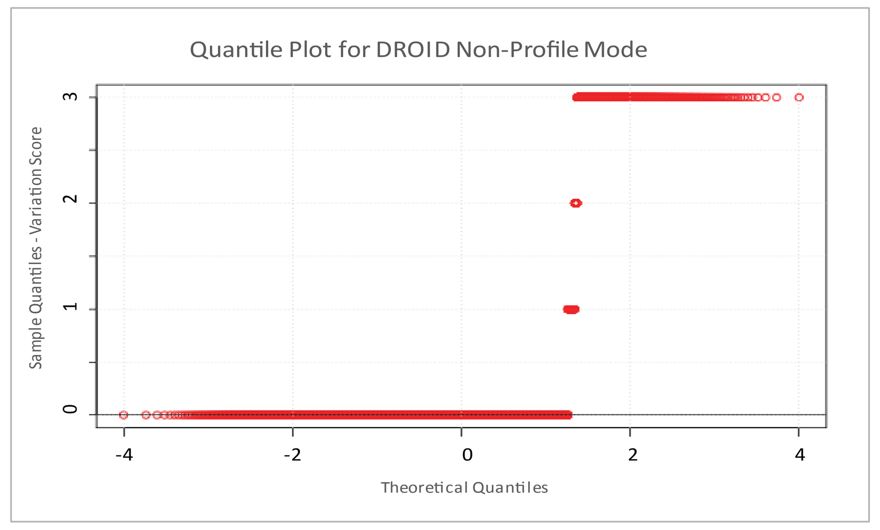
Siegfried is a signature-based file format tool that was originally developed in 2014 by Richard Lehane.

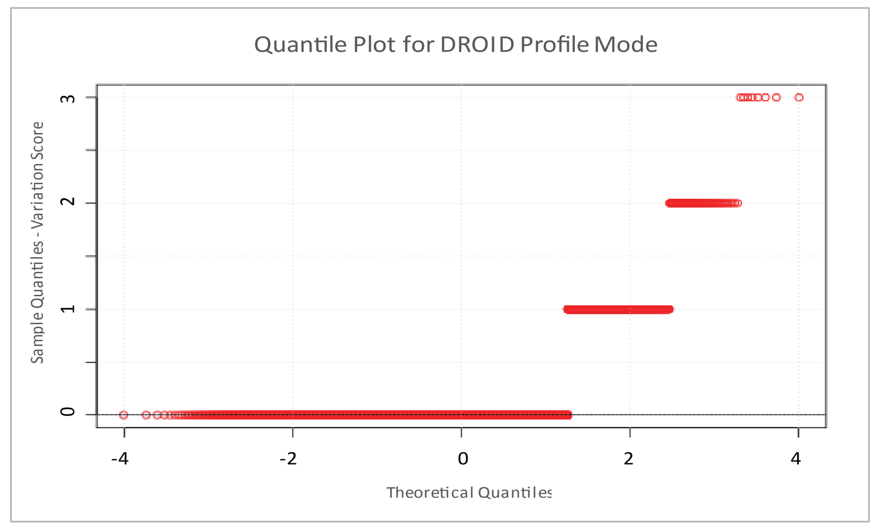
DROID, or Digital Record Object IDentification, was developed by the National Archives in the United Kingdom. The tool is run in both non-profile and profile modes.

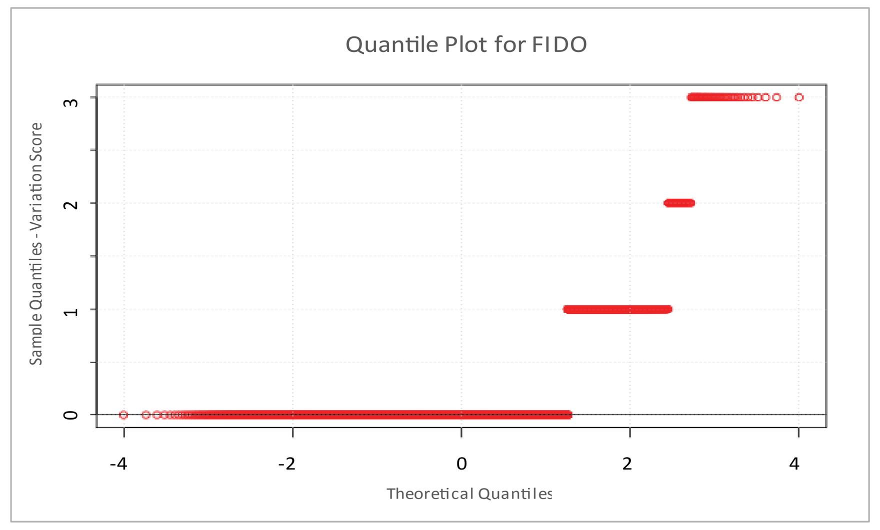
All four tools returned the same, consistent PUID values for 14,552 files. Distinct sets of inconsistent PUID values were returned for 1,678 or 10% of the files. Only 2 files were unknown by all four tools. The variation scores are graphed with quantile plots for each tool. Additionally, a line chart compares the mean, standard deviation, variance, and mode of the results from each tool. Lastly, the file extensions of files that produced the largest number of distinct sets are reported, e.g., files with a .m extension were identified with 11 different, distinct sets of PUIDs from the four tools.

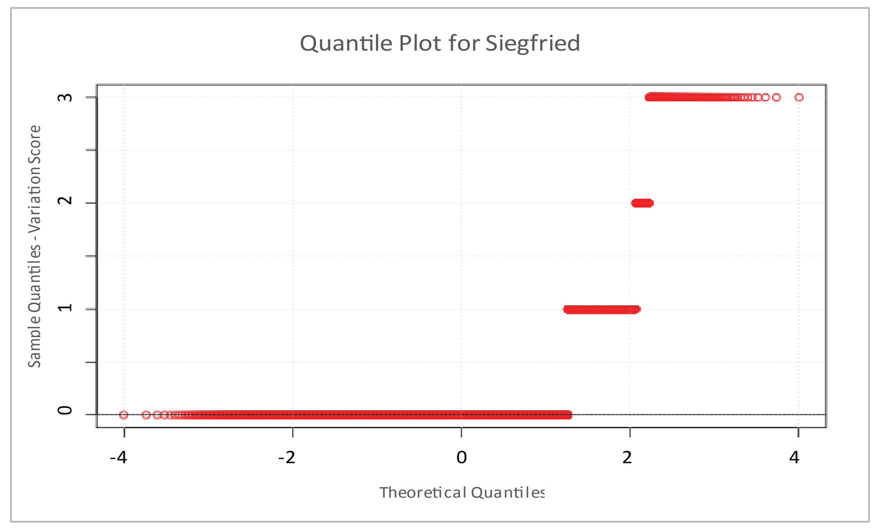


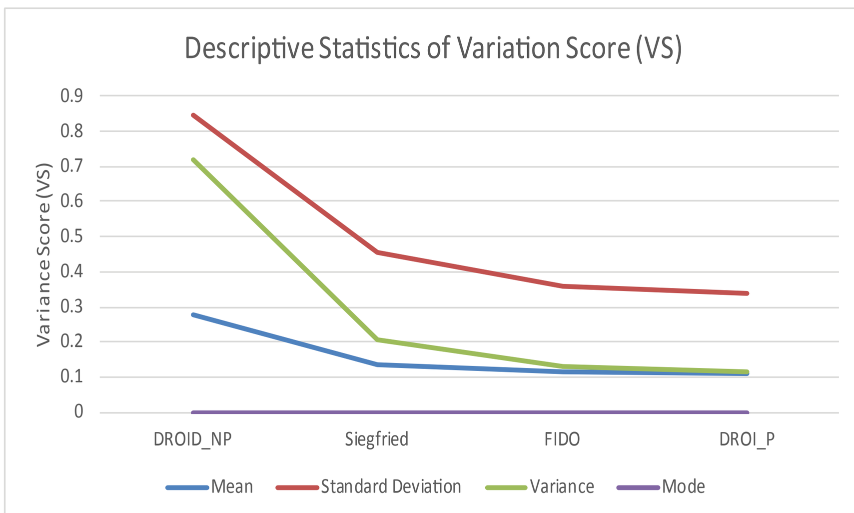


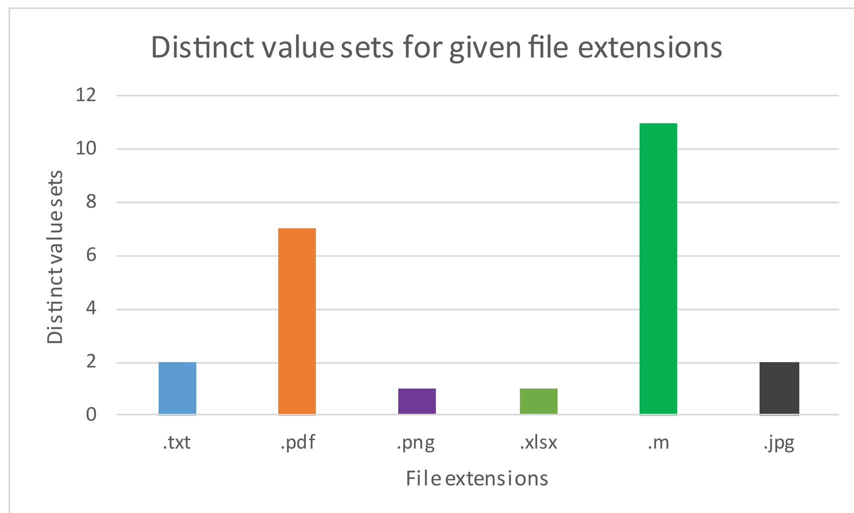












REFERENCES

Arms, C., & Fleischhauer, C. (2005). Sustainability of Digital Formats: Planning for Library of Congress Collections.

Abrams, S. L. (2005). Establishing a global digital format registry. *Library Trends, 54*(1), 125-143.

Brody, T., Carr, L., Hey, J., Brown, A., & Hitchcock, S. (2008). PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation, 2*(2), 3-19.

C. Dearborn, C., J. Barton, A., & A. Harmeyer, N. (2014). The Purdue University Research Repository: HUBzero customization for dataset publication and digital preservation. *OCLC Systems & Services, 30*(1), 15-27.

Lehane, R. (2014). Siegfried. Retrieved January 21, 2016, from http://www.itforarchivists.com/siegfried

Michael Witt (mwitt@purdue.edu)
Associate Professor of Library Science, Purdue University

Harish V. Kumaravel (hvalayap@purdue.edu)
Research Assistant, Purdue University

Carly Dearborn (cdearbor@purdue.edu)
Electronic Records Archivist, Purdue University

Yanqun Kuang (kuang5@purdue.edu)
Software Developer, Purdue University

D2C2 Distributed Data Curation Center

"Measuring the Accuracy of File Format Identification Tools"

Harish V. Kumaravel, Carly Dearborn, Michael Witt & Yanqun Kuang, Purdue University, West Lafayette, Indiana, USA