Digital Preservation File Format Policies of ARL Member Libraries: An Analysis

Kyle Rimkus, Thomas Padilla, Tracy Popp, Greer Martin

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

full paper to appear in March/April 2013 edition of d-Lib Magazine

Introduction

Whether overseeing institutional repositories, digital library collections, or digital preservation services, repository managers often establish file format policies intended to extend the longevity of collections under their care. While concerted efforts have been made in the library community to encourage common standards, digital preservation policies regularly vary from one digital library service to another.

This poster presents the findings of a study of file format policies at Association of Research Libraries (ARL) member institutions. It is intended to present the digital preservation community with an assessment of the level of trust currently placed in common file formats in digital library collections and institutional repositories.

The data show that file format policies have evolved little beyond the document and image digitization standards of traditional library reformatting programs, and that current approaches to file format policymaking must evolve to meet the challenges of research libraries' expanding digital repository services.

Data Model

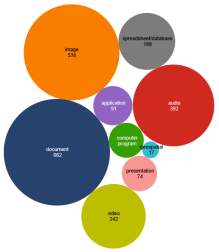
- Each **ARL Library** has zero or more instances of a **Repository or Digital Library Service**.
- Each **Repository or Digital Library Service** may enforce no more than one **File Format Policy**.
- Each **File Format Policy** must include one or more **File Format(s)**.
- Each **File Format** must belong to a **File Format Type** of the category **application**, **audio**, **computer program**, **geospatial**, **image**, **presentation**, **spreadsheet/database**, **text/document**, or **video**.
- Each **File Format** in a **File Format Policy** is supported at a **Confidence Level** of **High Confidence** or **Moderate Confidence**

ILLINOIS

Method

From October 2012-June 2013 the authors gathered preservation file format policy data from the 125 ARL member library websites and direct correspondence with digital repository and collection managers. Preservation file format support levels in these policies were normalized to moderate or high confidence and stored in a locally developed database.

The image below illustrates file format type distribution across all policies.



The authors determined Relative Confidence in a library's ability to preserve a given format by subtracting moderate confidence recommendations from high confidence recommendations, and dividing that result by the total number of recommendations.

(Hconfidence – Mconfidence) / Toccurence = Rconfidence

High Confidence is defined as:

- Any file format guaranteed functional preservation by virtue of the anticipated ability to preserve its content over time, to include formats designated for normalization or eventual migration to a secondary trusted file format.
- In lieu of a guarantee for functional preservation, any file format designated using language like "highly recommended," "high trust," or "high probability for digital preservation" in a file format policy that differentiates between high, moderate, and/or low levels of confidence.

Moderate Confidence is defined as:

- Any file format guaranteed bit-level but not functional preservation.
- Any file format designated using language like "weak" or "low trust" for digital preservation in a policy that differentiates between high, moderate, and/or low levels of confidence.
- Any file format listed as "accepted by" a repository or digital library service without any specific language designating the services implied by this acceptance.

Results

Fig. 1: Top 15 file formats by occurrence

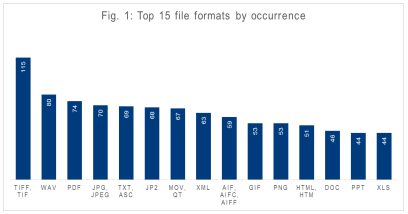


Fig. 2: File formats with positive relative confidence values

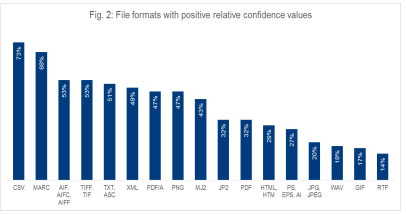


Fig. 3: Top 15 occurring file formats, with relative confidence value

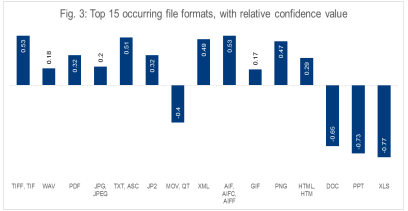
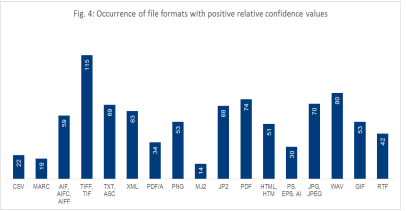


Fig. 4: Occurrence of file formats with positive relative confidence values



Conclusions

- Only 18 file formats have positive relative confidence values.
- Formats with positive relative confidence are largely born of library digitization programs and scholarly communication taking place on the web.
- Study indicates low level of confidence in ability to provide preservation services for most file formats.
- Future file format policies must expand the scope of relative confidence to ensure long lived access and reuse.

Acknowledgments

The authors wish to acknowledge the Research and Publication Committee of the University of Illinois at Urbana-Champaign Library, which provided support for the completion of this research.

Save File icon by iconoci
Book icon by Derrick Snider
Rock Collecting icon by Unknown Designer
Antenna icon by deadtype
Applause icon by Hum