Who's Afraid of File Format Obsolescence?

Developing a Data-Informed Research Agenda for Assessing File Format Endangerment

Heather Ryan • University of North Carolina at Chapel Hill

Curate Thyself Doctoral Symposium • Chapel Hill, NC • March 17, 2013

File formats pose a major risk to the long-term preservation of digital data and cultural heritage collections. Or do they? Digital preservation professionals and researchers cite impending file format obsolescence as a threat to the world's important digital information collections. For example, Pearson and Webb (2008) state that, "file format obsolescence is a major risk factor threatening the ongoing usefulness of digital information collections" (p.89). Others believe that file format obsolescence is not a threat. Most notably, Rosenthal (2010) states that, "there are no longer any plausible scenarios by which a format will ever go obsolete." The truth is that there is currently no substantial evidence that supports either case.

Through my research, I seek to fill this evidential gap. My goal is to develop a research road map to collect and analyze data on file formats that can be used to help researchers and digital collection managers make informed decisions regarding the viability of the digital file formats in their collections. Achieving this goal is a three-part process.

1 The first step is to establish baseline knowledge of file format **endangerment** levels, or the **degree to which a file format is "in danger" of becoming obsolete.**

2 The second step is to establish and test a method of **systematic data collection** based on reliable factors.

3 In the third step, data collection and **analysis** will produce **endangerment level ratings** that can be used to inform preservation action decisions.

1 Delphi Method

I will establish a baseline level of knowledge about file format endangerment by engaging file format experts in an online Delphi survey. The Delphi method was developed as a means to generate knowledge through expert polling and consensus. I will select experts to participate in this study based on their history in publishing literature on file format risk and by recommendations by other known experts in this area. For the first part of the survey, the experts will be asked log in to an online survey system where they will be prompted to rank endangerment levels of 25-50 file formats. They will also be asked to provide their rationale for their rankings. This process will repeat 2-3 times, wherein the participants will be able review each-others' choices and explanations, and may revise their answers based on the information provided by their fellow experts. Once the process has yielded agreement, I will create a baseline ranking of file format endangerment levels from the mean of the final expert rankings for each file format.

DRAFT Delphi Survey Tool Sample

Please rate each the following file formats' level of endangerment:

	Not endangered at all	Slightly Endangered	Quite Endangered	Nearly Obsolete	Currently Obsolete
PDF/A	○	○	○	○	○
AutoCAD DXF	○	○	○	○	○
QuarkXpress	○	○	○	○	○

Pearson, D. & Webb, C. (2008). Defining file format obsolescence: A risky journey. *The International Journal of Digital Curation, 3*(1), 89-106.
Rosenthal, D. (2010, November 24). *The half-life of digital formats*. Retrieved February 15, 2013 from http://blog.dshr.org/2010/11/half-life-of-digital-formats.html.

2 Factors for Data Collection

The second part of the survey will address factors for file format evaluation. I have discovered in the literature a dozen lists of factors for evaluating file formats. I will aggregate the factors used in the existing measures into a de-duplicated list. I will present the factors to the expert Delphi panel and they will rank each factor for usefulness and provide comments to explain the rationale for their choices. I will create the final list of factors based on the expert rankings.

System dependence
Security
Transparency
Complexity
Full open specification
Functionality
Long term stability
Self-Documentation
External Dependencies
Openness
Legal
Adoption
Error tolerance
Technical Protection Mechanism (DRM)
Robustness
Disclosure
Content fixity
Ease of handling

3 Future Research

Once I have created the list of factors, I will begin collecting data for each one. I will analyze the data using a number of methods established in epidemiology (Cumulative Sum algorithms), species endangerment monitoring (Population Viability Analysis), and anomaly detection (K-Nearest Neighbor, Naïve Bayes algorithms). I plan to evaluate each of these methods for effectiveness over time. I will triangulate expert opinion with statistical data to create endangerment level ratings.