ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Data Preservation

NSF

INSTITUTE of Museum and Library SERVICES

V1: Introduction

The nature and importance of data preservation (first pass)

Preservation is an important part of curation

A central objective of curation:

efficiently and reliably support the analysis of data,

and enable reuse over time

Despite:

- Physical deterioration or damage to storage devices
- Decay of digital representations (bit rot)
- Changes to file formats or encodings
- Changes to schemas and standards
- Changes to software tools and applications
- Loss or separation of critical contextual information or documentation
- Changes in practices and expectations *and so on*

(adapted from Giaretta, 2011)

A first definition

Digital preservation:

> "… the active management of digital content over time to ensure ongoing access." (US Library of Congress)

Obsolescence

File Format Obsolescence:

- Software upgrades don’t support legacy files
- Format is superseded by another or evolves in complexity
- File format is not compatible with current operating systems
- Company goes out of business or is bought by a competitor

Hardware/Media Obsolescence:

- New, faster computers and storage media
- Decrease in physical size of media (8in to 3.5in floppy disc)
- Reliability and fragility of media

Image from http://www.clir.org/pubs/reports/pub121/sec1.html

Physical Threats

Digital media and hardware are subject to numerous threats that can damage or destroy their readability:

- Environmental (temperature, humidity, light, dust, dirt)
  - Natural disaster (flood, earthquake)
  - Building failure (plumbing, electrical)
- Inevitable hardware failure
- Human error or improper handling
- Sabotage (theft, vandalism)

Context

Examples of metadata that must not be lost:

- File format:
  - Is this ASCII or EBCDIC?
  - Is this TEI version 2.1 or version 5.0?
- Data context
  - Who collected this observations? Where? When? How?
  - What do these attributes and values mean?
  - How were these values calculated?
- Processing
  - What software created this file?
  - What software can read and process (render, perform, etc.) this?