Automated Preservation Assessment of Heterogeneous Digital Collections

DLF Fall Forum
Charlottesville, VA
July 12–13, 2005

Keith Johnson

Product Manager, SDR　　keithj@stanford.edu

Preservation Workflows
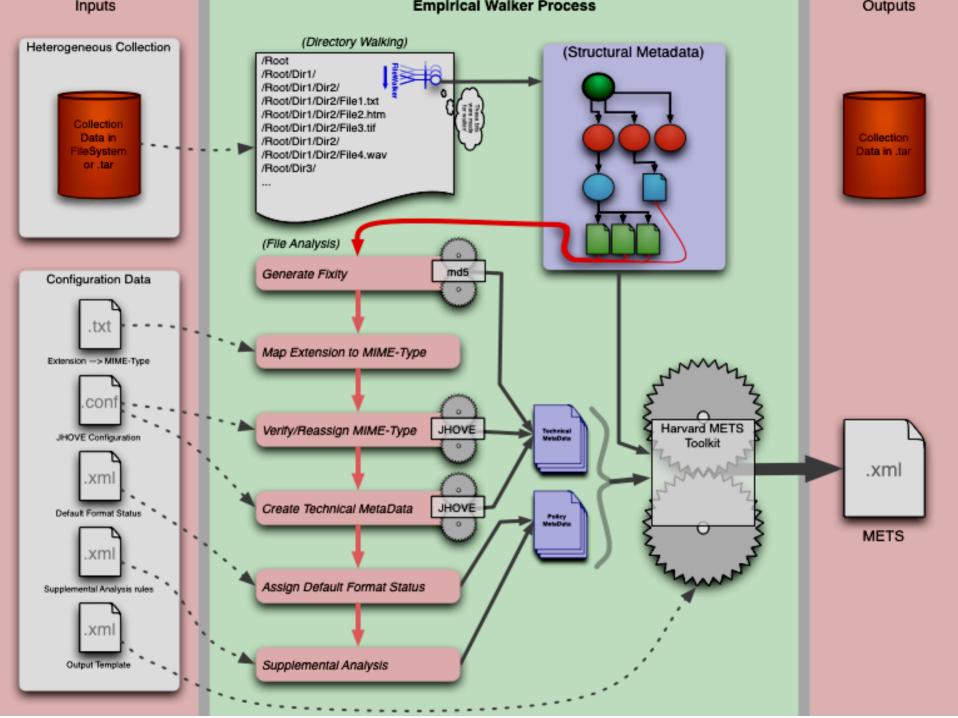
Prescriptive

"Library World"

Reactive

"Real World"

Quality to Policy

Default Format Score	Preservation Quality Status	Policy Status
0-1	High	Approved
2-3	Medium	Acceptable
4	Low	Minimal
5	Low	Unknown

SDR Preferred Formats

Plain text	ASCII
	UTF-8
Marked-up text	XML 1.0
Image	TIFF 5.0 and above (uncompressed)
Page-Viewer	PDF (any version)
Audio	WAVE (linear pulse code modulation)
Video	*to be determined*

Inputs

Heterogeneous Collection

Collection Data in FileSystem or .tar

Configuration Data

.txt

Extension —> MIME-Type

.conf

JHOVE Configuration

.xml

Default Format Status

.xml

Supplemental Analysis rules

.xml

Output Template

Empirical Walker Process

(Directory Walking)

/Root
/Root/Dir1/
/Root/Dir1/Dir2/
/Root/Dir1/Dir2/File1.txt
/Root/Dir1/Dir2/File2.htm
/Root/Dir1/Dir2/File3.tif
/Root/Dir1/Dir2/
/Root/Dir1/Dir2/File4.wav
/Root/Dir3/
...

(Structural Metadata)

(File Analysis)

Generate Fixity | md5

Map Extension to MIME-Type

Verify/Reassign MIME-Type | JHOVE

Create Technical MetaData | JHOVE

Assign Default Format Status

Supplemental Analysis

Technical MetaData

Policy MetaData

Harvard METS Toolkit

Outputs

Collection Data in .tar

.xml

METS

Policy Status Summary
(AIHT Files)

Policy Status	File Count	%
Preferred	25102	43.7
Approved	29929	52.1
Acceptable	1342	2.3
Minimal	0	0
Unknown	1077	1.9

Policy Status Summary
Post-Transform
(AIHT Files)

Policy Status	File Count	%
Preferred	41693	72.6
Approved	13338	23.2
Acceptable	1342	2.3
Minimal	0	0
Unknown	1077	1.9

The (Primary) Team

Richard Anderson, Stanford Digital Repository Software Developer

Catherine Aster, Digital Services Group Project Manager

Connie Brooks, Head, SULAIR Preservation

Hannah Frost, Media Preservation Librarian

Nancy Hoebelheinrich, SULAIR Metadata Coordinator

Keith Johnson, Stanford Digital Repository Project Manager,
  Project Manager, 05/2004–03/2005

Michael Keller, University Librarian; Director of Academic Information
  Resources; Publisher of HighWire Press; Publisher of Stanford
  University Press, Principal Investigator (PI)

Jerry Persons, SULAIR Chief Information Architect,
  Project Manager, 02/2004–04/2004