

Unit 13: Characterization of Digital Objects

DigCCurr Curriculum & Professional Institute - Digital Curation Exchange

Dr. Christopher (Cal) Lee & Dr. Helen R. Tibbo, UNC Chapel Hill SILS

Module Overview: Analysis & Characterization of Digital Objects

DigCCurr Curriculum (IMLS Grant #RE-05-06-0044 & #LG-05-09-0040)

School of Information and Library Science, University of North Carolina at Chapel Hill

1. Definition of Characterization

Identifying and documenting the properties of digital objects and information packages that are relevant to the ongoing curation, preservation, and use of the objects/packages. Includes identification of significant properties: 'properties of digital objects that affect their quality, usability, rendering, and behaviour' (InSPECT / Andrew Wilson).

2. Levels of Representation (Cal Lee)

- * Physical Object: Physical medium and magnetic/optical flux reversals (e.g., 3.5-inch disk).
- * Logical Object: Bitstream, sectors, file system structures, and encoded data.
- * Conceptual Object: Human-perceivable document, image, or interactive experience.

3. Core Stages of Characterization

- * Format Identification: Determining the presumptive file format and version using internal byte signatures (magic numbers), file extensions, and MIME types.
Key registries and tools: PRONOM (PUIDs), DROID, FITS, Unix file utility.
- * Validation: Determining whether an object conforms to the formal specification of its purported format (well-formedness vs. full validity; e.g., JHOVE).
- * Feature / Property Extraction: Extracting technical metadata embedded in headers and structures (dimensions, color space, compression, codecs, fonts, hashes).
Key tools: ExifTool, NLNZ Metadata Extractor, Apache Tika, FITS, MedialInfo.
- * Assessment & Significant Properties: Mapping technical properties to preservation action plans, migration pathways, and PREMIS representation metadata.

4. OAIS Representation Information & PREMIS Object Characteristics

In the ISO 14721 (OAIS) Reference Model, a Data Object (bit sequence) is interpreted using Representation Information (Structure Information + Semantic Information) to yield an Information Object meaningful to a Designated Community.

In PREMIS Data Dictionary for Preservation Metadata, <objectCharacteristics> captures:

- * <compositionLevel>: Layer of encoding/compression/encryption (0 = base bitstream).
- * <fixity>: Message digest algorithm (MD5, SHA-1, SHA-256) and value.
- * <size>: Size in bytes.
- * <format>: Format designation (name, version) and format registry key (e.g., PRONOM).
- * <creatingApplication>: Software name, version, and date created.
- * <significantProperties>: Type and value of properties deemed essential to preserve.
- * <objectCharacteristicsExtension>: Format-specific XML schemas (MIX, TEXTMD, AES57).

5. Characterization Tool Ecosystem

- * DROID (Digital Record Object Identification) - The National Archives (UK)
- * JHOVE (JSTOR/Harvard Object Validation Environment) - Open Planets Foundation
- * FITS (File Information Tool Set) - Harvard University Library wrapper combining DROID, ExifTool, FFIdent, JHOVE, National Library of New Zealand Metadata Extractor
- * BitCurator / Bulk Extractor / Fiwalk - DFXML generation from forensic disk images

6. Hands-On Lab Exercise: File Format Robustness & Inspection

1. Inspect file headers and magic numbers using a hex editor (e.g., 25 50 44 46 = %PDF).
2. Run DROID, JHOVE, and FITS against unmarked or mislabeled files.
3. Verify cryptographic hashes before and after controlled bit-level modifications.