METS: Technical Metadata for Text

DLF Forum, 22 January 2007

Jerome McDonough

New York University

jerome.mcdonough@nyu.edu

METS: Purpose of TextMD Schema

Provide sufficient information to:

- search text
- display text
- manipulate text
- migrate/preserve text

METS: TextMD Schema

- Encoding
  - Platform
  - Software
  - Agent
- Charset
- Byte order
- Line Break
- Language
- Font/Script
- Markup Basis
- Markup Language
- Processing Notes
- Print Requirements
- Viewing Requirements
- General Notes

TextMD: Encoding

Technical aspects of generating electronic text, including:

- Encoding platform – hardware used, including computer, scanner, etc.
- Encoding software – OCR, text editors, etc.
- Encoding agent – Those responsible for the process
- Quality Attribute – e.g. OCR recognition rate

TextMD: Characters

- Character set – employs controlled vocabulary (IANA Character Set Names)
- Byte order – if not clear from IANA charset name (big, middle, little endian)
- Line return – CR or CR/LF

TextMD: Writing

- Language – ISO 639-2 Codes
- Font / Script

TextMD: Markup

- Markup Basis – ‘meta’ markup language (GML, SGML, XML, etc.) and version information
- Markup Language – specific markup language (could be DTD or schema, or language such as LaTeX); may be a URI for DTD or schema, but not mandatory; should identify DTD or schema sufficiently to enable end user to identify a copy for validation

TextMD: Display

- Viewing Requirements – any special hardware or software necessary to display text to user
- Printing Requirements – any special hardware or software necessary to print text for user

TextMD: Notes

- Processing notes – any information regarding processing of text which may not have been covered in the encoding portion of textmd
- General notes – any other relevant information

TextMD: Example

```
<textMD><encoding>

        <encoding_platform>Dell Latitude C610</encoding_platform>

        <encoding_software>Windows 2000, SP2 </encoding_software>

        <encoding_agent role="MARKUP">J. McDonough</encoding_agent>

</encoding>

<charset>UTF-8</charset>
        <byte_order>little</byte_order>

<linebreak>CR/LF</linebreak>        <language>eng</language>

<markup_basis>XML Version 1.1</markup_basis>

<markup_language>http://dlib.nyu.edu/METS/textmd.xsd</markup_language>
```

METS: Other extension schema

- Video Technical Metadata – LC A/V available today; trying to work with SMPTE to develop a more extensive schema
- Intellectual Property Rights Metadata – Here there be dragons, and their attorneys

METS: Video Technical Metadata

- SMPTE Metadata Dictionary – basis for several other standards, including MPEG (www.smpte-ra.org)
- Aspect ratios (capture, presentation), gamma equation (capture, presentation), luma equation, colorimetry code, frame rate, lines (total frame, active frame, leading, trailing), luminance sampling rate, bits/pixel, compression, audio sampling, etc.

METS: Rights Language Metadata

- Existing Languages: XrML, ODRL
- In Development: Electronic Resource Management schema (thank you, Tim Jewell)
- Intellectual Property Rights for Intellectual Property Rights Languages: a painful patent primer

METS: Extension Schema Endorsement

- Endorsement: the METS editorial board may endorse schema, either to identify a preferred encoding practice in the case of competing schema for the same metadata set, or to promote the use of a particular metadata set as good practice
- Candidate schema for endorsement should be brought to the attention of the chair of the METS editorial board (jerome.mcdonough@nyu.edu)