Enabling Preservation by means of Open Source

dave rice
@dericed
#fosdem
2015-01-31
Demolition of Sony Magnetic Tape Plant - Dothan, Alabama

The demolition of the Sony Magnetic Tape Plant located in Dothan, Alabama
• Preservation of audiovisual material like the preservation of a species requires creation of copies.

• Preservation copies are generated to deter obsolescence risks.

• Archivists must negotiate challenges from increased collections size and constrained resources.

• One-on-one human-object interaction in preservation must be prioritized and selective rather than mandatory.

• The practicality of conservation-only approaches is weakened. Archives must be run differently than time capsules.
“How to move from one-to-one workflows to many-to-less workflows?”

“How can we preserve more than we could preserve before?”

“Is technology a replacement, a trusted co-worker? Is this a healthy relationship?”
Relax... your data will be backed up in 259,260 days.

FAIL

- Use your computer normally – backups happen in the background
- After the initial backup, incremental backups will be fast
- Your backup need not finish during your trial – you can test a restore anytime

Test your connection speed to Backblaze datacenter
Reformat to What?
Reformat to What?

Sustainability Factors
Disclosure
Adoption
Transparency
Self-Documentation
Metadata Capabilities
Impact of Patents
Reformat to What?

<table>
<thead>
<tr>
<th>Sustainability Factors</th>
<th>Cost Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disclosure</td>
<td>Cost of Software</td>
</tr>
<tr>
<td>Adoption</td>
<td>Cost of Hardware</td>
</tr>
<tr>
<td>Transparency</td>
<td>Storage Cost</td>
</tr>
<tr>
<td>Self-Documentation</td>
<td>Network Cost</td>
</tr>
<tr>
<td>Metadata Capabilities</td>
<td></td>
</tr>
<tr>
<td>Impact of Patents</td>
<td></td>
</tr>
</tbody>
</table>
Reformat to What?

**Sustainability Factors**
- Disclosure
- Adoption
- Transparency
- Self-Documentation
- Metadata Capabilities
- Impact of Patents

**Cost Factors**
- Cost of Software
- Cost of Hardware
- Storage Cost
- Network Cost

**Implementation Factors**
- Difficulty to Implement
- Complexity of Toolsets
- Availability to Tools: For transcoding, metadata, qc, etc
- Ease of Format Identification / Validation
LOC <3 JPEG2000!!!

Photo Credit: CC-BY-NC Sarah Ross Photography
SAMMA Inputs and Outputs

Tape Formats Supported
- Betacam, SP, SX, Digital Betacam
- U-matic small and large cassettes
- VHS, S-VHS, MIII, D3

Output Formats
- Video Tape, LTO Tape, DLT Tape
- SDI, Embedded audio or AES
- MPEG-2, H.264, IMX files
- Motion JPEG 2000 (lossless)
- Windows Media, QuickTime
You did it! Congratulations! World's best cup of coffee! Great job, everybody! It's great to meet you.
"As dissatisfying as it is, there is as yet no good, pat answer to digital video preservation, largely because the digital preservation world is so emergent. Efforts like FADGI’s [Federal Agencies Digitization Guidelines Initiative] MXF/JPEG 2000 work may help the digital preservation community to someday have a better answer to the 'which digital video format' question. For the moment the 'right answer' is actually more of a 'best set of questions to ask.'"

Uncompressed audiovisual streams in AVI, MXF, MOV?
I've been doing some experimenting with open-source tools to create JPEG2000-encoded MXF video files. I know this is something other people have been having trouble with, so I thought I'd share my findings.

I've been working on Mac OS X, but the tools should also be compatible with Linux and might work in Windows. I've found a few different paths to doing this. There isn't necessarily an ideal start-to-finish workflow yet, but the tools seem to be there to allow one to be scripted or programmed.

Method one: OpenDCP

- FFMpeg
- OpenJPEG 1.4 (http://www.openjpeg.org/)
- OpenDCP (http://code.google.com/p/opendcp/)

This workflow is designed around digital cinema package (DCP) production. Most of the open-source tools are based around DCP since they're being designed by indie filmmakers.

OpenDCP is a utility for creating DCPs. The AS-DCP library (http://www.cinecert.com/asdcplib/) and the mxlib library (http://sourceforge.net/projects/mxlib/) it's based on are possibilities for someone with more programming skills than me to build an archives-targeted tool.

The disadvantage to this route is that DCPs use separate files for audio and video, accompanied by an XML file describing the pair, which might make playback less convenient than a single AV MXF. The available tools
MKV / FFV1 / LPCM ?
# ffv1 implementations: Archivematica

<table>
<thead>
<tr>
<th>Media type</th>
<th>File formats</th>
<th>Preservation format(s)</th>
<th>Access format(s)</th>
<th>Normalization tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>AC3, AIFF, MP3, WAV, WMA</td>
<td>WAVE (LPCM)</td>
<td>MP3</td>
<td>FFmpeg</td>
</tr>
<tr>
<td>Email</td>
<td>PST</td>
<td>MBOX</td>
<td>MBOX</td>
<td>readpst</td>
</tr>
<tr>
<td>Email</td>
<td>Maildir**</td>
<td>Original format</td>
<td>MBOX</td>
<td>md2mb.py</td>
</tr>
<tr>
<td>Office Open XML</td>
<td>DOCX, PPTX, XLSX</td>
<td>Original format</td>
<td>PDF for PPTX</td>
<td>OpenOffice</td>
</tr>
<tr>
<td>Plain text</td>
<td>TXT</td>
<td>Original format</td>
<td>Original format</td>
<td>None</td>
</tr>
<tr>
<td>Portable Document Format</td>
<td>PDF</td>
<td>PDF/A</td>
<td>Original format</td>
<td>Ghostscript</td>
</tr>
<tr>
<td>Presentation files</td>
<td>PPT</td>
<td>Original format</td>
<td>PDF</td>
<td>OpenOffice</td>
</tr>
<tr>
<td>Raster images</td>
<td>BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA</td>
<td>Uncompressed TIFF</td>
<td>JPEG</td>
<td>ImageMagick</td>
</tr>
<tr>
<td>Raw camera files/Digital Negative format**</td>
<td>3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F</td>
<td>Original format</td>
<td>JPEG</td>
<td>ImageMagick/UF RAW</td>
</tr>
<tr>
<td>Spreadsheets</td>
<td>XLS</td>
<td>Original format</td>
<td>Original format</td>
<td>None</td>
</tr>
<tr>
<td>Vector images</td>
<td>AI, EPS, SVG</td>
<td>SVG</td>
<td>PDF</td>
<td>Inkscape</td>
</tr>
<tr>
<td>Video</td>
<td>AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV</td>
<td>FFV1/LPCM in MKV</td>
<td>MPEG-1</td>
<td>FFmpeg</td>
</tr>
</tbody>
</table>
ffv1 implementations: Österreichische Mediathek

DVA Profession

Video Digitization
DVA Profession is a complete solution for digitizing video for archival purposes. It manages the whole workflow, ranging from digitization to analysis, generating preview images and videos (MPEG), manual quality control, documentation of all process metadata and the final deposition of the files on a digital mass-storage. All steps of the workflow are designed and optimized for an economic operation and preparation for long-term archiving (for further information see "documentation"). This product is available under a Free Software License (GPLv3 - GNU General Public License) and can be downloaded here. Due to international participation, this site is kept in English.

Videodigitalisierung
ffv1 version 3

adds multithreaded encoding / decoding
adds mandatory embedded frame crc
self-descriptive
documentation efforts
Let’s Put Checksums inside Audiovisual Media

MP3
MPEG2
FLAC
FFV1 1.3
Checksum verification analogy

digipres

There's a problem
Checksum verification analogy

digipres

av archiving

There's a problem There's a problem
Checksum verification analogy

digipres

There’s a problem

av archiving

There’s a problem

av archiving with framemd5

There’s a problem
<table>
<thead>
<tr>
<th>#tb</th>
<th>0: 1001/30000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 0, 0, 0, 1, 1228800, 39cc3d1589ea601881abddd07fc8a32b</td>
<td></td>
</tr>
<tr>
<td>0, 1, 1, 1, 1228800, bbd6d0d757017ae0f75a505e589f5dc</td>
<td></td>
</tr>
<tr>
<td>0, 2, 2, 1, 1228800, 94b7c46fb0d456209fca3206bb3f20f2</td>
<td></td>
</tr>
<tr>
<td>0, 4, 4, 1, 1228800, b4a2e57b784164716a7f4f5c4b1fa22e</td>
<td></td>
</tr>
<tr>
<td>0, 5, 5, 1, 1228800, 1d16a3d98d9b2667419ad84814d9c74e</td>
<td></td>
</tr>
<tr>
<td>0, 6, 6, 1, 1228800, 28bdf2224aac89bb445d3e55755792bb</td>
<td></td>
</tr>
<tr>
<td>0, 7, 7, 1, 1228800, 4b15f8b8f27e94aa515c76b83fa566ef</td>
<td></td>
</tr>
<tr>
<td>0, 8, 8, 1, 1228800, a544793f6bb30d97ccf43549e74c7748</td>
<td></td>
</tr>
<tr>
<td>0, 10, 10, 1, 1228800, f222da8e7be0dfba8f6f70fbbaf82f47</td>
<td></td>
</tr>
<tr>
<td>0, 11, 11, 1, 1228800, bc42b9f41147d4db582ce91f3ce641b3</td>
<td></td>
</tr>
<tr>
<td>0, 12, 12, 1, 1228800, dabc1bd60575626e3f46c6726b632e8be</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#tb</th>
<th>0: 1001/30000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 0, 0, 0, 1, 1228800, 39cc3d1589ea601881abddd07fc8a32b</td>
<td></td>
</tr>
<tr>
<td>0, 1, 1, 1, 1228800, bbd6d0d757017ae0f75a505e589f5dc</td>
<td></td>
</tr>
<tr>
<td>0, 2, 2, 1, 1228800, 94b7c46fb0d456209fca3206bb3f20f2</td>
<td></td>
</tr>
<tr>
<td>0, 4, 4, 1, 1228800, b4a2e57b784164716a7f4f5c4b1fa22e</td>
<td></td>
</tr>
<tr>
<td>0, 5, 5, 1, 1228800, 1d16a3d98d9b2667419ad84814d9c74e</td>
<td></td>
</tr>
<tr>
<td>0, 6, 6, 1, 1228800, 35f67d5a7dd9eab6938a38b56d785648</td>
<td></td>
</tr>
<tr>
<td>0, 7, 7, 1, 1228800, 4b15f8b8f27e94aa515c76b83fa566ef</td>
<td></td>
</tr>
<tr>
<td>0, 8, 8, 1, 1228800, a544793f6bb30d97ccf43549e74c7748</td>
<td></td>
</tr>
<tr>
<td>0, 10, 10, 1, 1228800, f222da8e7be0dfba8f6f70fbbaf82f47</td>
<td></td>
</tr>
<tr>
<td>0, 11, 11, 1, 1228800, bc42b9f41147d4db582ce91f3ce641b3</td>
<td></td>
</tr>
<tr>
<td>0, 12, 12, 1, 1228800, dabc1bd60575626e3f46c6726b632e8be</td>
<td></td>
</tr>
</tbody>
</table>
FLAC are internally verified by md5 per stream and crc per audio frame

ffmpeg -v warning -i sine.flac -f null -
[flac @ 0x7fd0cb05e800] CRC error at PTS 96768
“[ffv1 @ 0x7f9855046e00] CRC mismatch FC686A4F! frame 215”
Significant Characteristics

Frame Size
Frame Rate
Color Space: Color Matrix
Luma Range: Broadcast or Full
Aspect Ratio
Interlacement
Duration / Timeline
Audio Channel Configuration / Arrangement
static const char *j2ki_field_rate_companies[] = { "Canopus", "SAMMA" };

static int64_t klv_decode_ber_length(AVIOContext *pb)
{
    uint64_t size = avio_r8(pb);
}

static int mxf_add_timecode_metadata(AVDictionary **pm, const char *key, AVTimecode *tc)
{
    AVDictionary *d = mxf_add_dictionary(pm, key, "TimeCode", tc);
    av_dict_free(d);
    return 0;
}

static int mxf_parse_structural_metadata(MXFContext *mxf)
{
    AVFormatContext *s = mxf->fc;
    MXFPackage *material_package = NULL;
    MXFPackage *temp_package = NULL;
    int i, j, k, ret;

    case AV_CODEC_ID_JPEG2000:
    if (descriptor->frame_layout == SegmentedFrame ||
        descriptor->frame_layout == SeparateFields) {
        st->codec->time_base = st->time_base;
        st->time_base = (AVRational) { st->time_base.num, st->time_base.den * 2 };
        int l, field_rate = 0;
        AVDictionaryEntry *entry = av_dict_get(s->metadata, "company_name", NULL, 0);
        if (entry) {
            for (l = 0; l < FF_ARRAY_ELEMS(j2ki_field_rate_companies); l++) {
                if (av_stristr(entry->value, j2ki_field_rate_companies[l])) {
                    av_log(s, AV_LOG_INFO, "J2ki sample rate will be interpreted as field rate for company: %s
                    field_rate = 1;
                    break;
                }
            }
        }
    }
    return 0;
}
Sustaining Consistent Presentation
PREFORMA, FUTURE MEMORY STANDARDS

Overview
Memory institutions are facing increasing transfers of electronic documents and other media content for long term preservation. Preservation models are often inspired by ISO 14721:2003, known as “the OAIS model”, where transfers and preservation are built on information packages containing both data and metadata.

Data are normally stored in specific file formats for documents, images, sound, video etc. that are produced by software from different vendors. Even if the transferred files are in standard formats, the implementation of standards cannot be guaranteed. The software implementing standards for the production of the electronic files is not in control neither by the institutions that produce them nor by the memory institutions. Conformance tests of transfers are done, but are not totally reliable. This poses problems in long-term preservation. Data objects meant for preservation, passing through an uncontrolled generative process, can jeopardise the whole preservation exercise.

The overall intention of PREFORMA project (PREservation FORMAAts for culture information/e-archives) is to research critical factors in the quality of standard implementation in order to establish a long-term sustainable ecosystem around developed tools with a variety of stakeholder groups. The tools should be innovative and provide a reference implementation of the most common file format standards for the assessment of the collections to be archived and for the
Enabling Preservation by means of Open Source

dave rice
@dericed
#fosdem
2015-01-31