Digital Shelf Life

Building Files to Last

Joel Eaton

Technical Support Officer
for Sound, JISC Digital Media

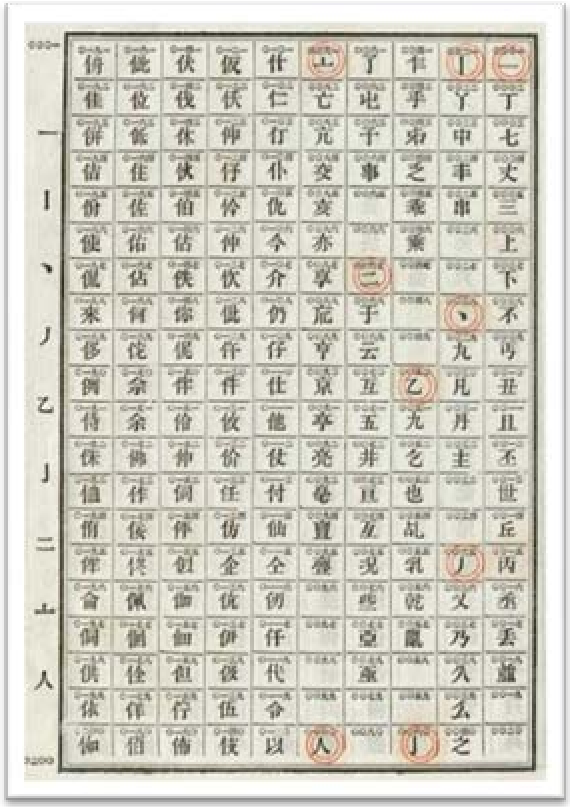
info@jiscdigitalmedia.ac.uk

JISC Digital Media
Still images, moving images and sound advice

What’s the risk?

Chinese Telegraphy pre-2002

Decode Morse code into sequence of digits and chop these into quadruplets. Then decode these one by one with reference to the ‘restricted’ operators’ manual.



JISC Digital Media
Still images, moving images and sound advice

Real World Scenario #1

- Higher education collection of videos
- Research council funded digitisation
- Digital raw data is ‘safe’: duplicated in x2 locations, error checked
- But no on-going funding meant no updating
- 3 years after completion…

UNRECOGNISED FILE TYPE!_

Real World Scenario #2

- Library collection of mixed media
- Images, video, sound and e-books all held in readable formats
- But library management system built by (ex) student
- Catalogue data in unique format, upon import into a new system…

UNRECOGNISED FILE TYPE!_

So who recommends file types?

- **Submission** guidelines for repositories
- Policies created for long term **preservation**
- Format registries

JISC Digital Media

Still images, moving images and sound advice

Submission guidelines for repositories

American Geophysical Union

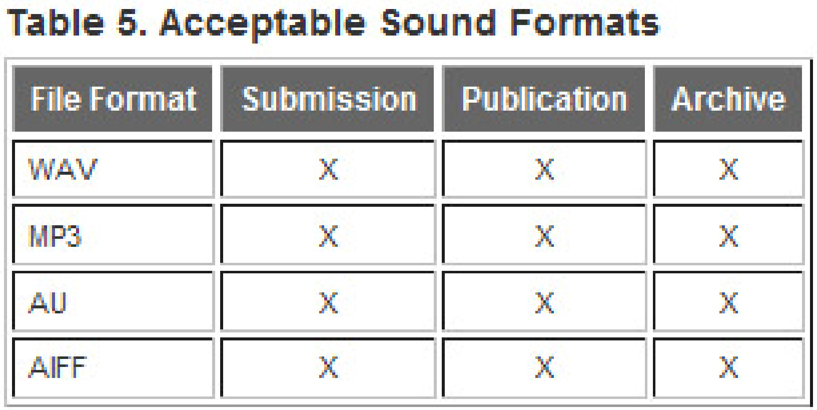


Table 5. Acceptable Sound Formats

File Format	Submission	Publication	Archive
WAV	X	X	X
MP3	X	X	X
AU	X	X	X
AIFF	X	X	X

http://www.agu.org/pubs/authors/manuscript_tools/journals/formats.sht

Policies created for long term preservation

Arts & Humanities Data Service

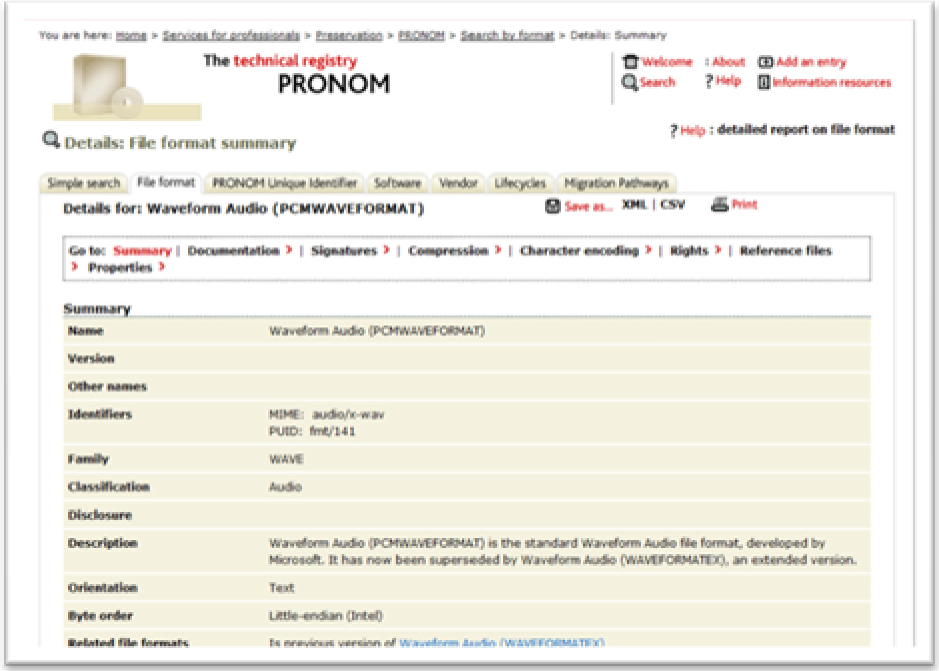
Preferred Audio Formats: WAV, AIFF

http://www.ukoln.ac.uk/web-focus/papers/ichim05/html/

JISC Digital Media
Still images, moving images and sound advice

Format registries

PRONOM


You are here: Home > Services for professionals > Preservation > PRONOM > Search by format > Details: Summary
The technical registry
PRONOM
Welcome
: About
Add an entry
Search
? Help
Information resources
? Help : detailed report on file format
Simple search
File format
PRONOM Unique Identifier
Software
Vendor
Lifecycles
Migration Pathways
Save as... XML | CSV
Print
Go to: Summary | Documentation › | Signatures › | Compression › | Character encoding › | Rights › | Reference files › Properties ›

Details: File format summary

Details for: Waveform Audio (PCMWAVEFORMAT)

Summary

Name	Waveform Audio (PCMWAVEFORMAT)
Version	
Other names	
Identifiers	MIME: audio/x-wav PUID: fmt/141
Family	WAVE
Classification	Audio
Disclosure	
Description	Waveform Audio (PCMWAVEFORMAT) is the standard Waveform Audio file format, developed by Microsoft. It has now been superseded by Waveform Audio (WAVEFORMATEX), an extended version.
Orientation	Text
Byte order	Little-endian (Intel)
Related file formats	Is previous version of Waveform Audio (WAVEFORMATEX)

http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

JISC Digital Media
Still images, moving images and sound advice

Which type is the right type?

“It’s not possible to recommend a definitive list of formats... it is possible to establish selection criteria which can be used to help repositories”

DPC File Formats for Preservation

The five selection criteria:

1. Widespread adoption
2. A lack of technological dependencies
3. The disclosure of specification
4. Transparency i.e. ‘identifiability’
5. Ability to embed metadata

JISC Digital Media
Still images, moving images and sound advice

1.Widespread adoption

- Is the format heavily used (a de facto standard)?
- Is the format being used heavily and in the *correct sector*?*
- Vendors have an agenda, how do we *know* the format is popular?*
- Proprietary often beats open source

* Look for public sector surveys

JISC Digital Media
Still images, moving images and sound advice

2.Lack of technological dependencies

- Formats should be compatible with many software and hardware systems
- Complex files (e.g. content with wrapper format) may add dependencies
- Should be made by many different manufacturers
- Opensource often beats proprietary

3.Disclosure within public realm

- Even poorest formats are usable (but may be uneconomical to recover) if code is in the public domain
- Heavily customised code can go down with a sinking ship and take your data with it!
- Again, opensource often trumps proprietary

4.Transparency of format & content

- Formats should have good representation information to allow easy identification
- Again, wrapper/content formats can be problematic

5.Ability to embed metadata

- Without context files become inaccessible
- Embedded metadata offers extra protection against a centralised system failure
- Metadata not always text-based, not always human readable
- Embedded metadata need not comprehensive, can be used with a centralised system

JISC Digital Media
Still images, moving images and sound advice

Other criteria might be

- Can it be repurposed?
- Is the format simple to use?
- Is the format evolving or is it stable?
- Can the format be ‘locked’ via DRM?
- Is it expensive to use?

Q. After looking at the criteria we've selected ***format_x***. Will it last?

A. No, all formats will become obsolete and will need to change over time*

*But should still 'perform' in the same way

Amazing ‘performing’ data

- Files become akin to a musical score (rather than a gramophone record)
- The *Performance Model*: ISO 15489
- Preservation strategy of the National Archives of Australia
- Files should *convey the essence* of a digital record

How does your data perform?

- Files ‘do’ lots of things, which do you really care about?
- Define your *significant properties*
- Ensure these are maintained, regardless of current or future file types

JISC Digital Media
Still images, moving images and sound advice

So which type is the right type?

“align with a clear **preservation strategy** that articulates the purpose of the repository and the needs of its community”

[formats] “must be appropriate to the needs of the repository”

DPC File Formats for Preservation