Introduction to Research Data Management

File formats and metadata

LIBRARIES
TEXAS A&M UNIVERSITY

Workshops

1. Build an overview
2. Collect and document data
3. Store digital data
4. Work with data
5. **Share and preserve data**
6. Plan ahead

Introduction

Focus: Understanding file formats for data, and taking advantage of your documentation to create metadata that describe your data files.

The goal is to ensure your data are usable when shared with others.

Discussion

What kind of file formats you are likely to work with in your research?

File formats

File formats are a way of encoding information within a computer file, the format specifies how bits are used to encode information.

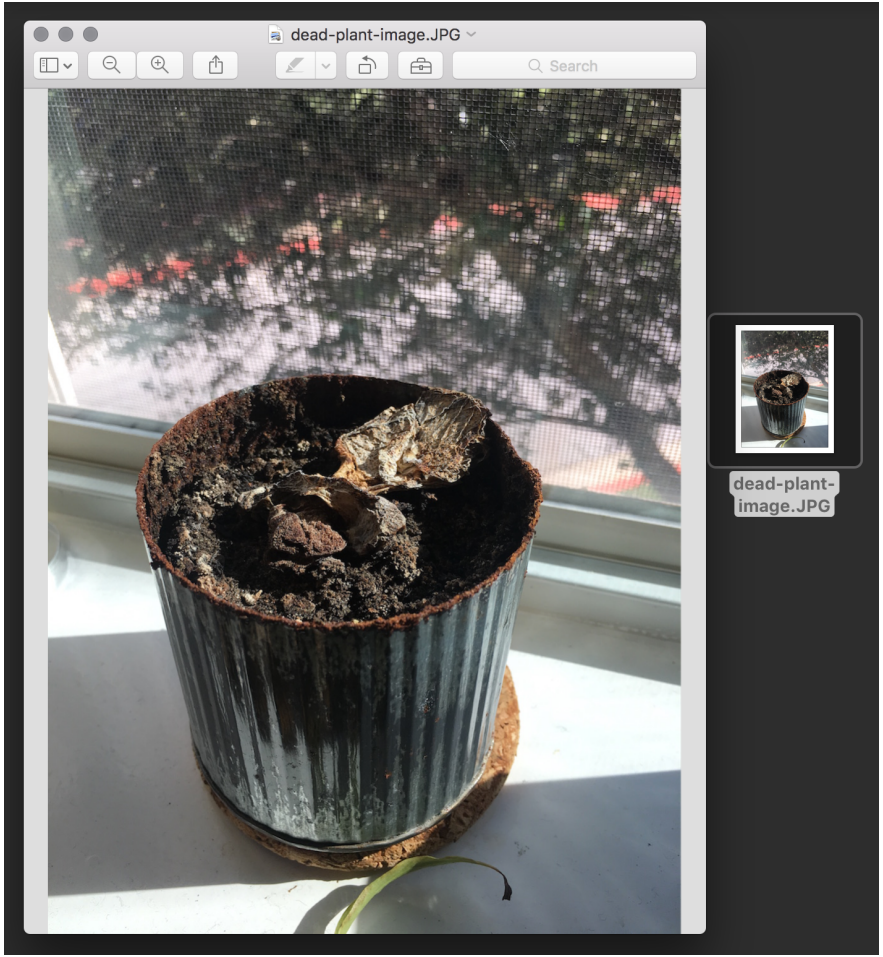
Two types of file formats:

- Binary
- Text

Files with binary encodings

Can only be read by applicable software, and can contain formatting information, images, sounds, compressed versions of other files.

May use proprietary standards.


dead-plant-image.JPG
Search
dead-plant-image.JPG
dead-plant-image.JPG — Edited
dead-plant-image.JPG

Files with character encodings

Text files use character encoding standards to make them machine-readable.

All text characters are encoded but many standards are in use, often depending on software and country.

ASCII and UTF-8 text encoding standards

Use ASCII or Unicode UTF-8 character encoding for “plain text” file formats, TXT, CSV, HTML, XML.

ASCII: Used to represent the alphabetic, numeric, and punctuation characters commonly used in English.

UTF-8: Backward compatible with ASCII. Capable of encoding all valid code points in Unicode, to cover the characters of most alphabets.

File formats for data

From the data lifecycle perspective, think about:

Interoperability: data files are usable with different software tools.

Preservation: data files can be accessed 10 or more years later.

Features of formats that last

1. In common usage by the research community.
2. Non-proprietary.
3. Open and documented standards.
4. Uncompressed (space permitting).
5. Use standard character encodings.

Recommended formats

Content	File formats
Text	PDF/A, HTML, XML, TXT
Tabular data (spreadsheets and databases)	XML, CSV
Numbers and statistics	TXT, DTA, POR, SAS, SAV
Geospatial	SHP, DBF, GeoTIFF, NetCDF
Audio	WAVE, AIFF, MP3, MXF
Images	TIFF, JPG, JP2, PDF, PNG, GIF, BMP
Moving Images	MOV, MPEG, AVI, MXF
Web Archive	WARC
Containers	TAR, GZIP, ZIP

Alternatives

Discouraged Format	Alternative Format
Excel (.xls, .xlsx)	Comma Separated Values (.csv)
Word (.doc, .docx)	plain text (.txt), or if formatting is needed,
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Photoshop (.psd)	TIFF (.tif, .tiff)
Quicktime (.mov)	MPEG-4 (.mp4)

Tips for converting file formats

1. Go from proprietary or software-based to non-proprietary or open formats.
2. Use to standard character encodings.
3. Beware of lossiness and corruption.

Containers

ZIP: De facto standard (lossless) compression format that is used on Windows, Mac, and Linux platforms

TAR: Commonly used in Unix and Linux to bundle a set of files into one.

Metadata

Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

Metadata is often called “data about data” or “information about information.”

Discussion

Have you created or used metadata?

What is it good for?

Metadata for data

From the data lifecycle perspective, information for:

Discoverability: data files can be located and identified correctly.

Accessibility: content within data files can be interpreted, assessed, and used.

Distinguishing documentation and metadata

Documentation

- Created while working on a project.
- Can be informal.
- May pertain to many different levels (project, datasets, data files, variables and values).
- May provide general context.

Metadata

- Often created upon “publication” of an object.
- Formally describes a particular object (can be a data file or dataset).
- Formatted into a record and structured according to particular schemas.
- May derive from the documentation.

Types of metadata

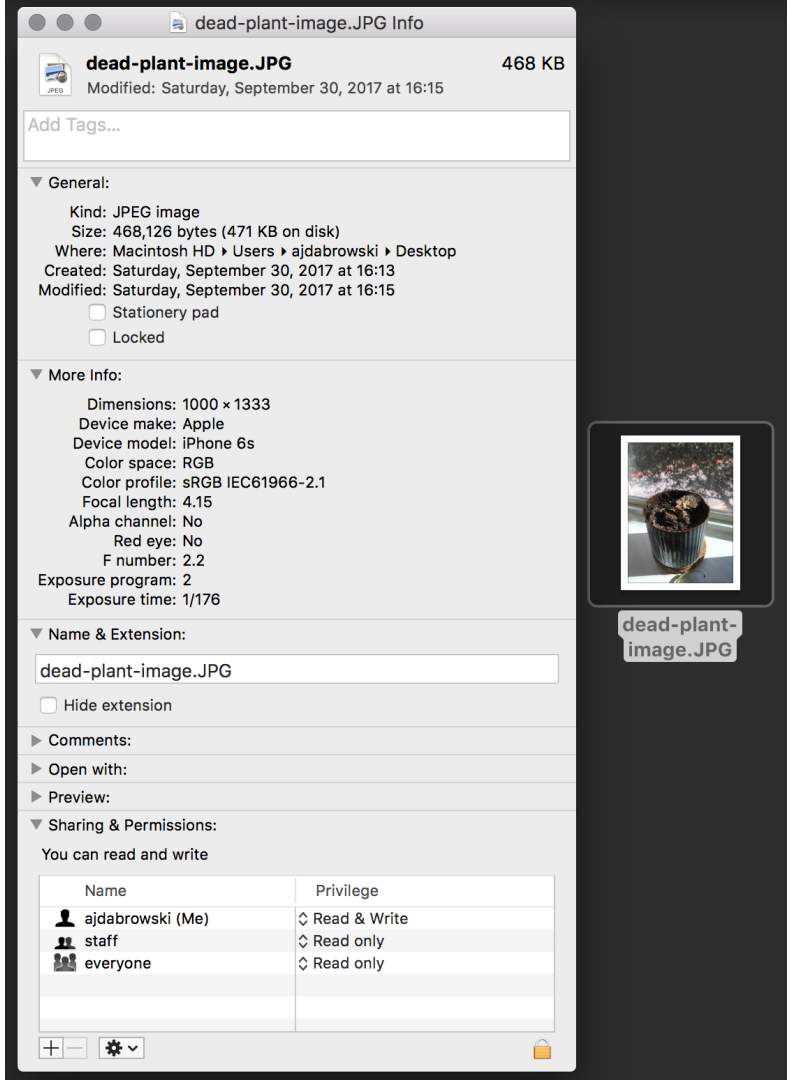
- **Descriptive**: Describes the object and gives the basic facts.
- **Structural**: Describes the structure of an object including its components and how they are related.
- **Administrative**: Contains information about the management of the object, e.g. terms of use, required software, provenance (history), and file integrity checks.

Creating metadata

Machine generated.

User generated.

- Copy-pasted from documentation.
- Created on the spot.



Metadata records

Core descriptive metadata:

- Title
- Creator
- Identifier
- Subject
- Dates

Metadata standards

Guide the collection and structure of metadata so that data is collected, described, structured, and referred to consistently.

Examples of metadata standards

Find disciplinary standards: http://www.dcc.ac.uk/resources/metadata-standards

Discipline	Standard
Biology	Darwin Core
Ecology	EML - Ecological Metadata Language
Earth Sciences	AgMES - Agricultural Metadata Element Set
Physical Sciences	CIF - Crystallographic Information Framework
Social Sciences & Humanities	DDI - Data Documentation Initiative
General Research Data	DataCite Metadata Schema
General Research Data	Dublin Core

Creating metadata records

- Controlled vocabularies: lists of predefined terms that ensure consistency of use, and help to disambiguate similar concepts.

- Technical standards: ensure that the units such as date and time are entered consistently amongst different researchers.

Examples of controlled vocabularies

- ERIC Thesaurus for education terms (http://www.eric.ed.gov/ERICWebPortal/resources/html/thesaurus/about_thesaurus.html)
- IEE INSPEC Thesaurus of the Scientific and Technical terms (http://www.theiet.org/resources/inspec/products/aids/index.cfm)
- Centre for Agricultural Bioscience international's CAB Thesaurus (http://www.cabi.org/cabthesaurus/mtwdk.exe?yi=home)
- Medical Subject Headings (MeSH) (http://www.nlm.nih.gov/mesh/)
- Library of Congress Subject Headings (LCSH) (http://authorities.loc.gov/)

Date and time standard

ISO 8601

- Year: YYYY (e.g. 1997)
- Year and month: YYYY-MM (e.g. 1997-07)
- Complete date: YYYY-MM-DD (e.g. 1997-07-16)
- Complete date plus hours and minutes: YYYY-MM-DDThh:mmTZD
    - (e.g. 1997-07-16T19:20+01:00)
- Complete date plus hours, minutes and seconds: YYYY-MM-DDThh:mm:ssTZD
    - (e.g. 1997-07-16T19:20:30+01:00)

Tips for metadata

- Consistent data entry is important.
- Avoid extraneous punctuation, it can create retrieval issues.
- Avoid most abbreviations.
- Use templates and macros when possible.
- Extract metadata from your documentation.
- Keep a reference to elements, technical standards, and controlled vocabularies you use in your project.
- Always use an established metadata standard.

Conclusion

- Identified appropriate file formats for sharing data.
- Reviewed basics of metadata standards

References and resources

- Abrams, Stephen. "File Formats" [PDF](http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/file-formats)
- DataONE. "Document and Store Data Using Stable File Formats" [Website](http://www.dataone.org/best-practices/document-and-store-data-using-stable-file-formats)
- Library of Congress. "Sustainability of digital formats" [Website](https://www.loc.gov/preservation/digital/formats/index.shtml)
- NISO. "Understanding Metadata: What is metadata and what is it for?" [PDF](http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf)