
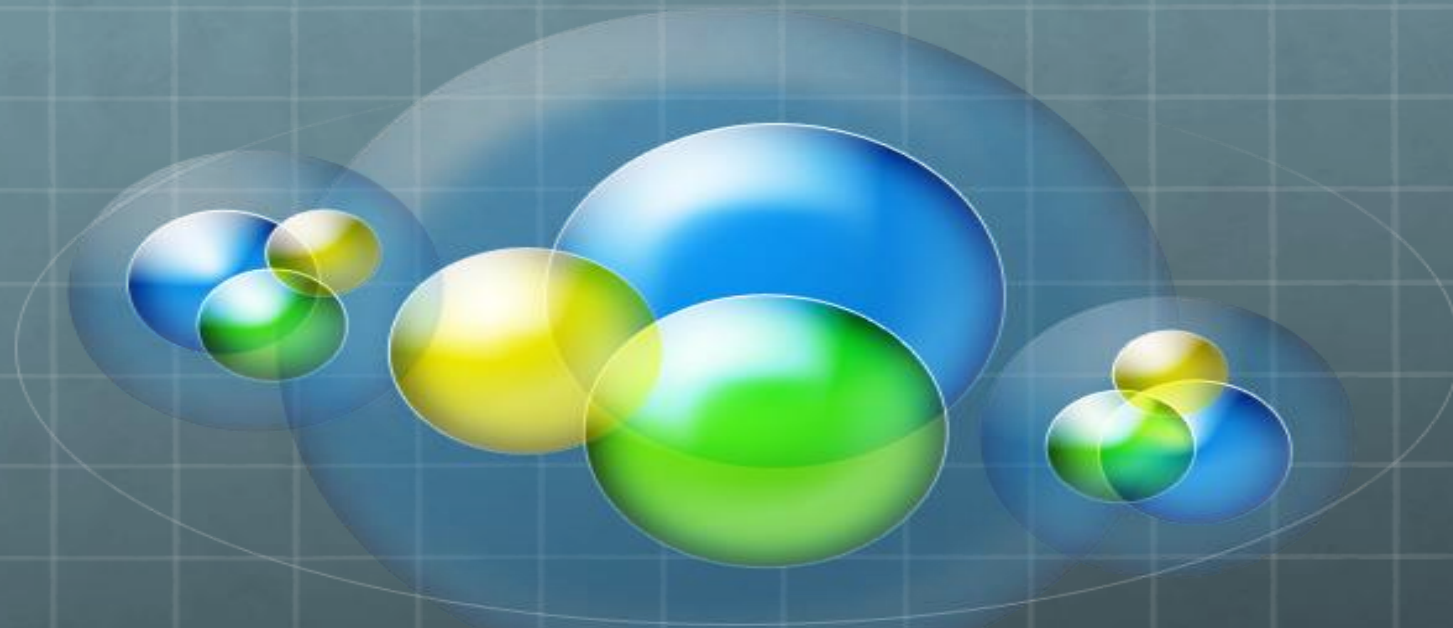


# Data Management and Metadata

-  We'll be doing an activity at the end of this session, so go ahead and download the data set to your computer:

<http://www.zemkat.org/RDSC/pets.php>



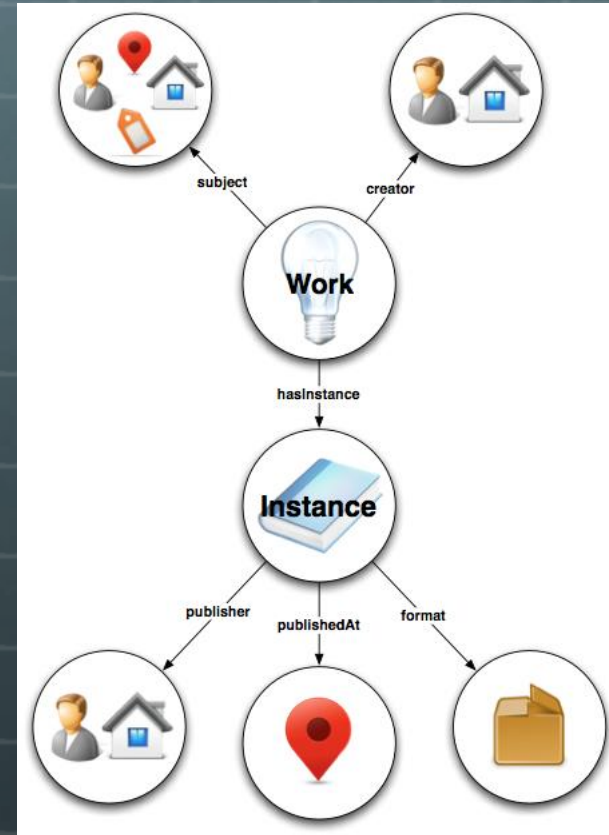
# Data Management and Metadata

Kathryn Lybarger  
@zemkat

University of Kentucky Libraries  
February 23, 2017

# What is metadata?

- 🌐 “data about data”
- 🌐 Many formats and schemas
- 🌐 Found in many places
- 🌐 Can be brief or very detailed
- 🌐 Varying structure
- 🌐 Many functions of metadata



# Cataloging

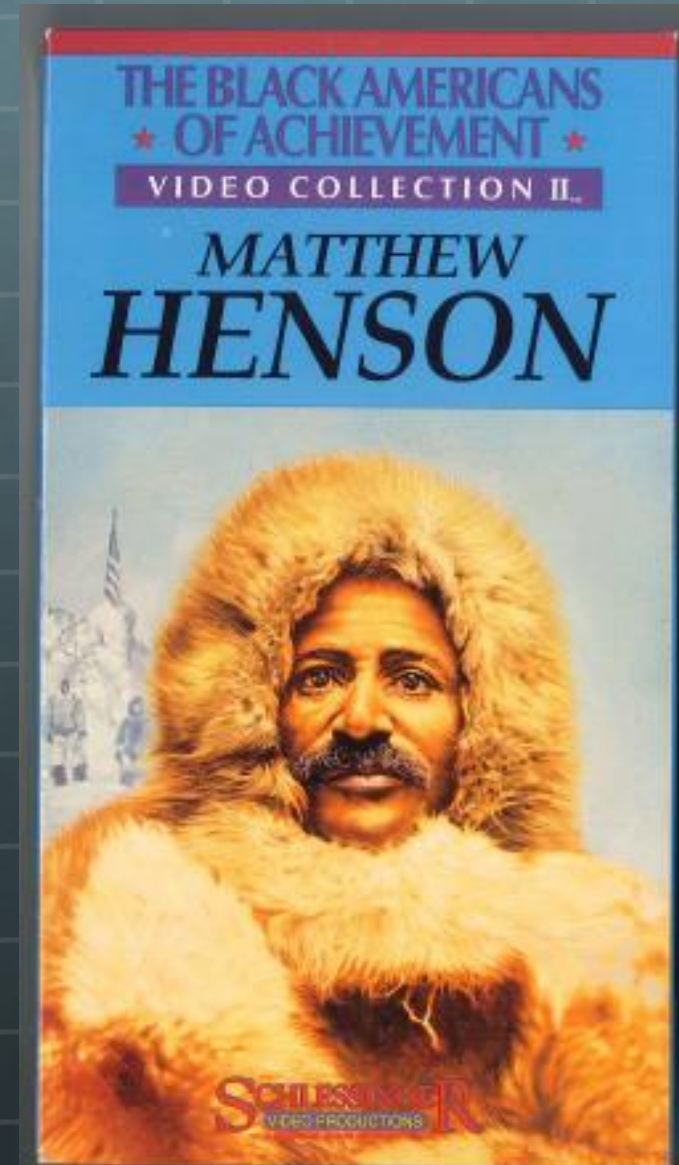
- Describing bibliographic data
- Usually in a library context
- Content Standards
  - RDA: Resource Description and Access
  - AACR2: Anglo-American Cataloging Rules
- File formats
  - MARC – MACHine Readable Cataloging
  - BIBFRAME



By Dr. Marcus Gossler (Own work)  
[GFDL (<http://www.gnu.org/copyleft/fdl.html>)  
or CC-BY-SA-3.0 via Wikimedia Commons

# Other metadata

```
<fileSec>
  ▼<fileGrp ID="pageFileGrp1">
    ▼<file ADMID="mixmasterFile1 premismasterFi
      <Flocat LOCTYPE="OTHER" OTHERLOCTYPE="fi
    </file>
    ▼<file ADMID="mixserviceFile1 premisservice
      <Flocat LOCTYPE="OTHER" OTHERLOCTYPE="fi
    </file>
    ▼<file ADMID="premisotherDerivativeFile1" I
      <Flocat LOCTYPE="OTHER" OTHERLOCTYPE="fi
    </file>
    ▼<file ADMID="premisocrFile1" ID="ocrFile1"
      <Flocat LOCTYPE="OTHER" OTHERLOCTYPE="fi
    </file>
  </fileGrp>
  ▼<fileGrp ID="pageFileGrp2">
    ▼<file ADMID="mixmasterFile2 premismasterFi
      <Flocat LOCTYPE="OTHER" OTHERLOCTYPE="fi
    </file>
```



# Metadata by any other name...

- 🌐 The cover
- 🌐 Documentation
- 🌐 Annotation
- 🌐 Markup
- 🌐 File headers
- 🌐 Finding aids



Created by Alice Noir  
from Noun Project

# General schema: Simple Dublin Core

 Title

 Contributor

 Source

 Creator

 Date

 Language

 Subject

 Type

 Relation

 Description

 Format

 Coverage

 Publisher

 Identifier

 Rights

# General schema: Qualified Dublin Core

- 🌐 **Date**
- 🌐 **Created**
- 🌐 **Valid**
- 🌐 **Available**
- 🌐 **Issued**
- 🌐 **Modified**



Created by Michal Beno  
from Noun Project






# Discipline-specific schema

## Libraries

-  EAD – Encoded Archival Description
-  DACS – Describing Archives: a Content Standard
-  MODS - Metadata Object Description Schema

## Other disciplines:

-  Darwin Core (Life Sciences)
-  DDI - Data Documentation Initiative (Arts & Humanities)
-  ISO 19115 (Geography)



Created by Bemar Novalyi  
from Noun Project

# Data should be FAIR

- 🌐 **F**indable – good metadata, indexed somewhere
- 🌐 **A**ccessible – data is retrievable once you've found it
- 🌐 **I**nteroperable – follows common metadata standards
- 🌐 **R**e-usable – richly described, follows community standards
  
- 🌐 The FAIR Guiding Principles for scientific data management and stewardship
  - 🌐 <http://www.nature.com/articles/sdata201618>



Love Your Data  
Week  
#LYD17

Feb. 13-17

# For the future!

**METADATA IS A  
LOVE NOTE  
TO THE FUTURE**

Quote by Jason Scott (@textfiles)

By cea + from The Netherlands (Metadata is a love note to the future) [CC BY 2.0 via Wikimedia Commons]

# ... the very near future.

- 🌐 **Eagleson's Law of Programming:**  
Any code of your own that you haven't looked at for six or more months, might as well have been written by someone else.
- 🌐 **For data, how long will you remember:**
  - 🌐 **What's in which file?**
  - 🌐 **When and how it was collected?**
  - 🌐 **How to use the data?**




Created by Andrew Forrester  
from Noun Project

# Metadata can be simple


- 🌐 File naming / directory structure
- 🌐 Files have a “creation date”
- 🌐 Photos know when and where they were taken
- 🌐 Filenames have an file format extension
- 🌐 Spreadsheet columns have labels
  
- 🌐 This is all good *metadata*, but you should have a plan

# Descriptive metadata







## Discovery

-  Allows you to search the metadata, and find the data you're looking for

## Identification

-  If you found the right metadata, would you know it was what you were looking for?

## Examples:

-  Title
-  Creator
-  Keywords or tags
-  Identifier
-  Geospatial coverage
-  Date

# Date formats

- Which is best?
  - February 23, 2017
  - Thursday, February 23, 2017
  - 2/23/17
  - 2017-02-23
  - 23 February 2017



# ISO-8601



- 🌐 2017-02-23 (YYYY-MM-DD)
- 🌐 Allows:
  - 🌐 Standardized recording of dates, times (weeks!)
  - 🌐 Sorting just works
    - 🌐 Even if some dates have different granularity
- 🌐 International standard
  - 🌐 Also common!



# Dates in file names / headers

```
!<8f>Exif^@^@II*^@^H^@^@^@M^@^@  
^@^@^@À^L^@^@^A^A^D^@^A^@^@^@<90  
^O^A^B^@^H^@^@^@a^@^@^@P^A^B^@^  
^@^@^@R^A^C^@^A^@^@^@A^@^@^@Z  
^@^@^@Â^@^@^@[^A^E^@^A^@^@^@Ê^@  
C^@^A^@^@^@B^@^@^@1^A^B^@^L^@^@  
@2^A^B^@^T^@^@^@p^@^@^@S^B^C^@^  
A^@^@^@i<87>^D^@^A^@^@^@ò^@^@^@%  
^A^@^@^@: ^S^@^@L^S^@^@SAMSUNG^@S  
GH-I747H^@^@^@A^@^@^@H^@^@^@A^  
7UCDLK3^@^@2013:05:03 15:18:41^@  
<82>^E^@^A^@^@^@`^B^@^@<9d><82>^  
@^@h^B^@^@"<88>^C^@^A^@^@^@C^@^  
^C^@^A^@^@^@}^@^@^@^@<90>^G^@^D^
```



20130503\_151841.jpg

JPEG image - 1.7 MB

Created Friday, May 3, 2013 at 3:18 PM

Modified Friday, May 3, 2013 at 3:18 PM

Last opened Friday, May 3, 2013 at 3:18 PM

Dimensions 3264 × 2448

[Add Tags...](#)

# Excel's date format



Created by Sean Maldjian  
from Noun Project

- It's very insistent that you use it
  - Maybe okay if this data will never leave this file
- Looks weird:
  - 11/4/11
  - Aug-16
- Internally: # of days since 1900 (or 1904 on Mac!)
- Exports / converts poorly
- I use ISO-8601 in a text field

# Structural metadata

- 🌐 Indicates how different parts of the data set relate
- 🌐 Examples
  - 🌐 Relationships to other file sets
  - 🌐 Relationships between different files
  - 🌐 Same data in different file types



Created by Gregor Cresnar  
from Noun Project

# File formats

## For different purposes

- For image data:
  - TIFF – original capture
  - JPEG – for web display
  - JPEG 2000 – for multi-res display
  - PDF – for easy distribution
  - XML – OCR text

## For preservation / use

- For scientific data
  - DAT – internal format from instrument
  - Excel – export for general use
  - CSV – most portable
  - README (.txt) – documentation for that file

# Administrative metadata

- 🌐 Broad category, including:
  - 🌐 Technical metadata
  - 🌐 Preservation metadata
  - 🌐 Rights metadata
  - 🌐 ...



Created by Gregor Cresnar  
from Noun Project

## A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#\$\$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file





Type: Ph.D Thesis Modified: too many times

Copyright: Jorge Cham


www.phdcomics.com

# Version control


## Keep track of:

-  When files changed
-  What content changed
-  Who changed them
-  Why






## Keep file system tidy:

-  No need to include “002” or “FINAL” in file names

## Backups:

-  Retrieve / restore to any previous versions

## Examples:

-  Git (GitHub)
-  Subversion
-  Mercurial
-  (“Track changes” / “Past versions”)
-  (“Shadow copies”)

## Hopefully repository-provided!

# Technical metadata

## Helps to:

- Decode
- Render
- Interpret

## Examples:

- File format
- Is it compressed?
- Has any processing been done?
- How was data gathered?
- What equipment was used?
- Using what settings?



Created by Mert Güler  
from Noun Project



# Know your tools / procedure

When cataloging a book in RDA, we measure the dimensions and record something like:

24 cm

This means:

Height is between 23 and 24 cm

Width is between 12 and 24 cm

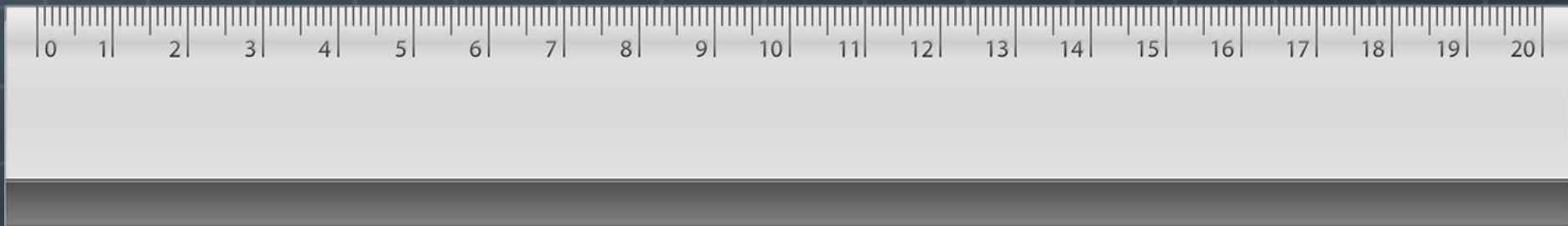
Tools:

Standard ruler (cm)

Procedure:

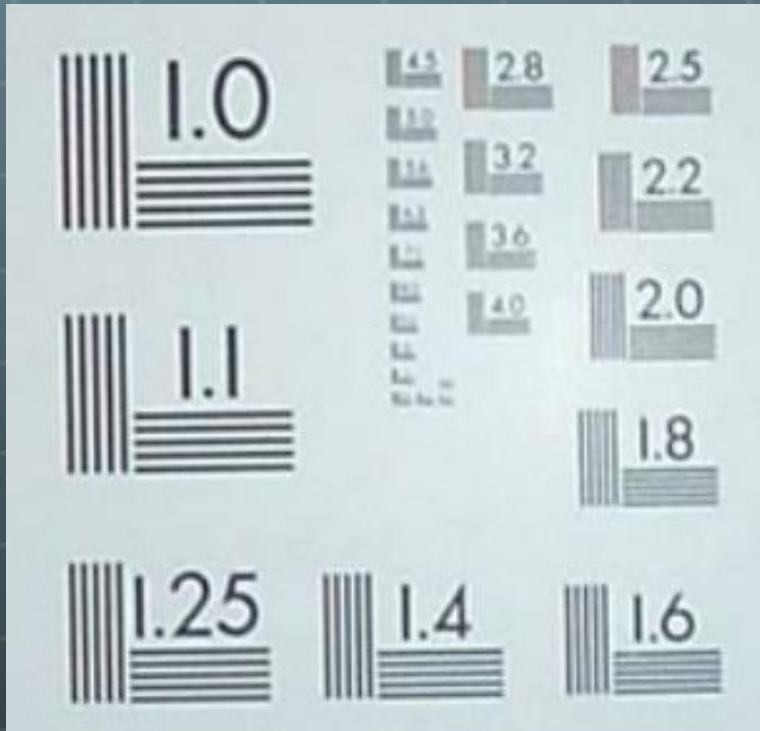
Measure height in centimeters, round up

Only include width if greater than height, or less than half of height

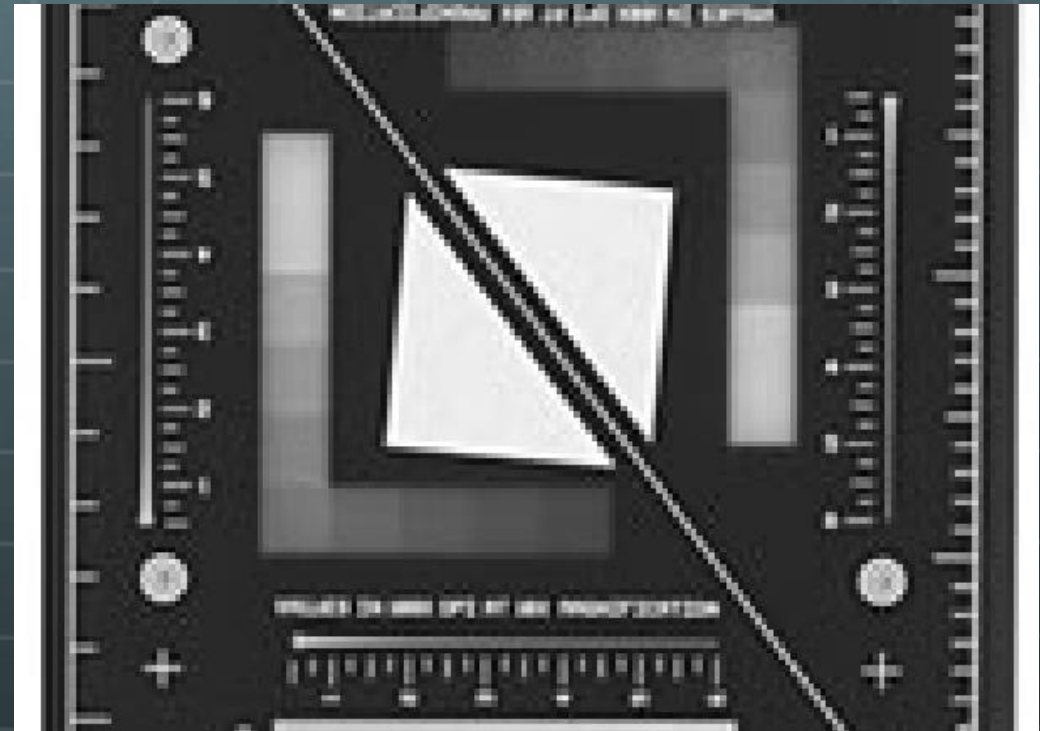


# Technical targets

For microfilm



For digital capture



# Identifying art forgery

- NOVA challenged teams to identify art forgeries based on digital images of them
- Hypothesis:
  - Copies have higher density of brush strokes
  - (“they tried too hard”)
  - Higher contrast throughout image
- Results:
  - Success!



NOVA's "Catching a Copy" and "Art Authentication"

# But...

- 🌐 Images of some paintings were taken with a newer camera
- 🌐 Higher resolution registered more detail / contrast in brush strokes
- 🌐 This technical metadata was not taken into account
- 🌐 Results of data analysis were flawed



# Preservation metadata

- Helps with long-term management of data
- Is this the same data set?
  - Are all the files here?
  - Same versions?
  - Has data rot occurred?
  - Has a file been truncated or otherwise corrupted?
  - Did it survive a transfer?
- Common format: PREMIS
  - XML, auto-generated
  - Checksums



Created by Viktor Vorobyev  
from Noun Project

# ISBN (has a check digit)

9780747544593

- Add up digits in **odd** positions
- Add to that digits in **even** positions x 3
- Divide by 10, subtract remainder from 10 to get **check digit**:
  - $9+8+7+7+4+5 + 3(7+0+4+5+4+9) = 127$
  - $127 / 10$  has remainder 7
  - $10 - 7 = 3$
- (if remainder is zero, check digit is zero)

# Checksums

- Run an algorithm on your file to create a much smaller file (checksum)
- Can be used to detect if two files are the same
  - Did the file download correctly / completely?
  - Has the file changed over time?
- If the file changes (gets replaced, truncated, etc.) the checksum will be different
- Examples:
  - md5sum, sha1sum, cksum

# Rights metadata

- Intellectual property rights attached to data
  - (How) can you access, use, or re-use the data?
- Copyright?
- Are there confidentiality issues?
- A license, like Creative Commons?





# Open Data

- 🌐 “Open data is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike.”–Open Data Handbook
- 🌐 CCo has no restrictions
- 🌐 CC-BY – appropriate attribution required
- 🌐 CC-BY-SA – share-alike – copies or adaptations  
must be under the same license as the original.



Created by Arthur Shlain  
from Noun Project

# What metadata should you use?

- 🌐 If you're going to deposit your data into a repository, do they have requirements?
- 🌐 Does your discipline have a common standard?
- 🌐 What would be helpful in searching / using your data?

# What does metadata for data really look like? (it depends)

Photo,Lat,Long,Lat deg,Lat min,Lat sec,Long deg,Long min,Long sec,File name

a,42.96971667,131.9017,42,58,10.98,131,54,6.12,DSC02235

b,44.77659333,132.0156881,44,46,35.736,132,0,56.477,DSC01920

c,44.92947972,131.6624031,44,55,46.127,131,39,44.651,DSC02028

d,44.92931167,131.66225,44,55,45.522,131,39,44.1,DSC02022

e,44.891675,131.588665,44,53,30.03,131,35,19.194,DSC01984

f,48.94773139,136.2745717,48,56,51.833,136,16,28.458,DSC01062

[https://doi.org/10.13012/B2IDB-4084515\\_V1](https://doi.org/10.13012/B2IDB-4084515_V1)

# Directory structure / file naming

- Common practice
- Easy to browse
- More difficult to search
- Fragile
- Still a good idea?
  - Yes, but can do more



# README files

- 🌐 **Text file accompanying your data**
  - 🌐 Plain text (not MS Word)
- 🌐 **How many?**
  - 🌐 One README for the whole data set
  - 🌐 One README per file or directory
- 🌐 **For tabular data, definitions of columns**
- 🌐 **How has the data been processed?**
- 🌐 **Follow a consistent structure**
  - 🌐 May be a standard

```
sketch_feb3a.ino  ReadMe.adoc  ▼
:Author: zemkat
:Email: arduino@zemkat.org
>Date: 04/02/2017
:Revision: version#
:License: Public Domain

= Project: {Project}

Describe your project

== Step 1: Installation
Please describe the steps to install this project.

For example:

1. Open this file
2. Edit as you like
```

# Database / Spreadsheet

- 🌐 Useful if you have lots of data files
- 🌐 More restrictive in format
- 🌐 More easily searchable
- 🌐 (May need its own README)

	A	B	C
	Filename	Date	Lab
	107315024.csv	2017-01-23	Green
	600900099.csv	2017-01-24	Green
	795096600.csv	2017-01-26	Green
	2246626095.csv	2017-01-29	Green
	4294967295.csv	2017-01-31	Green
	salazar.csv	2017-01-23	Red
	casper.csv	2017-01-24	Red
	markov.csv	2017-01-26	Red
	cheese.csv	2017-01-29	Red
	helsinki.csv	2017-01-31	Red

# Activity: Library pet data set

<http://www.zemkat.org/RDSC/pets.php>

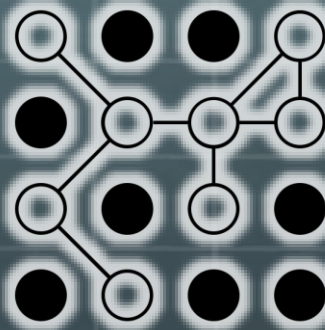
- 🌐 Download files
- 🌐 Gather into groups
- 🌐 Create some metadata
- 🌐 I'll be asking questions like:
  - 🌐 Find a picture of one of Kathryn's pets
  - 🌐 How many pictures have dogs?
  - 🌐 How many pictures taken before 2013?



# What does your metadata look like?



Created by Harsha Rai  
from Noun Project



Created by Viktor Vorobyev  
from Noun Project





# Activity questions

 Question 1:

**Find a cat. What's its name?**

**Find a non-mammal. What kind of animal is it?**

# Activity questions

 Question 2:

**Find one of Cindy's pets. What's its name?**

**How many pets' names start with P?**

# Activity questions

## Question 3:

**Find a picture taken in 2013. Who's in it?**

**How many pictures taken after May 2016?**

# Activity questions

 Question 4:

**How many pets live indoors?**

**How many pets greet you at the door?**

# What problems did you have?

- 🌐 Would you have done something differently when gathering the data?
- 🌐 Having answered the questions, would you do something different assigning metadata?

# Now swap!



Created by Delwar Hossain  
from Noun Project

🌐 Exchange metadata  
with a neighbor, or  
download mine:

🌐 [http://www.zemkat.org/  
LibraryPets/](http://www.zemkat.org/LibraryPets/)

# Activity questions

 Question 1:

**Find a bird. What's its name?**

**Find a picture with two animals. Who are they?**

# Activity questions

 Question 2:

**Find one of Mary's pets. What's its name?**

**How many pets' names start with R?**



# Activity questions

## Question 3:

**Find a picture taken in 2014. Who's in it?**

**How many pictures taken after October 2016?**

# What problems did you have?

 Did you have better luck with your own?

 Do you think they had better luck with yours?

# References

- NISO Understanding Metadata Primer
  - [http://www.niso.org/publications/press/understanding\\_metadata/](http://www.niso.org/publications/press/understanding_metadata/)
- The FAIR Guiding Principles for scientific data management and stewardship
  - <http://www.nature.com/articles/sdata201618>
- Open Data (Creative Commons)
  - <https://creativecommons.org/about/program-areas/open-data/>
- ReadMe Guidance (Dryad at NCSU)
  - <http://datadryad.org/pages/readme>

**Thank you for coming!**

**Please fill out the online evaluation form, linked here:**

**<http://www.zemkat.org/RDSC/pets.php>**