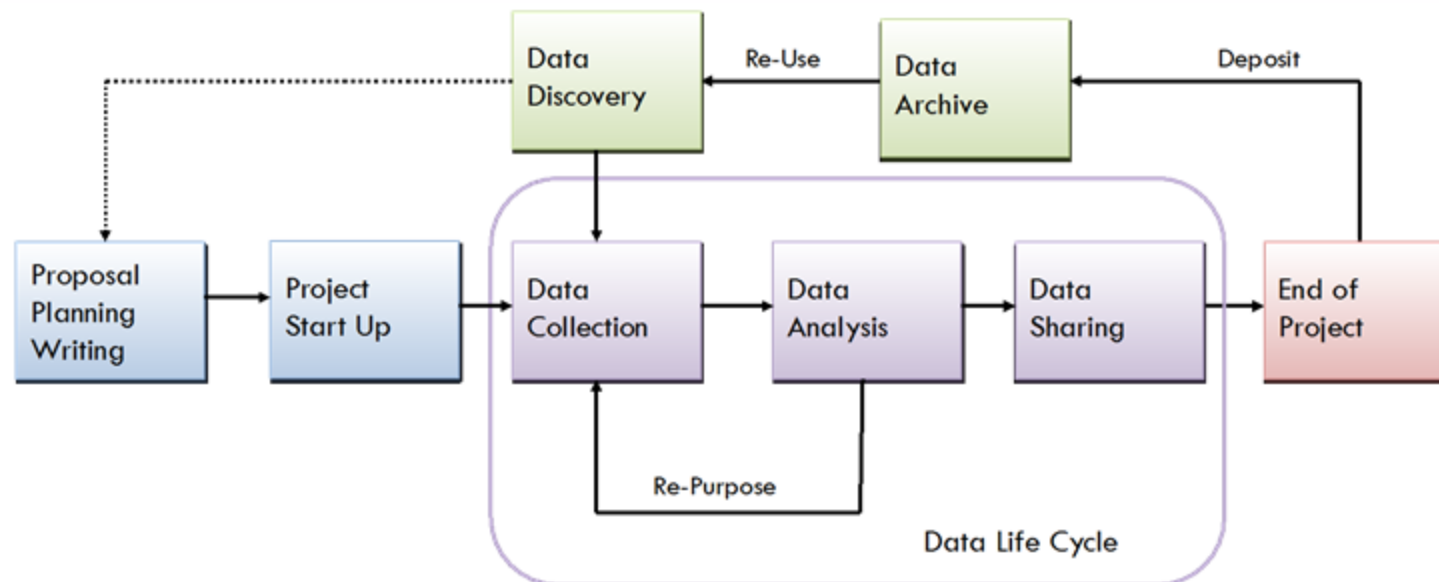# Documentation and Metadata

Anne Gaynor

amgaynor@virginia.edu

Sherry Lake

shLake@virginia.edu

Website for Today's Workshop

http://data.library.virginia.edu/data-management/plan/metadata/metadata-workshop/
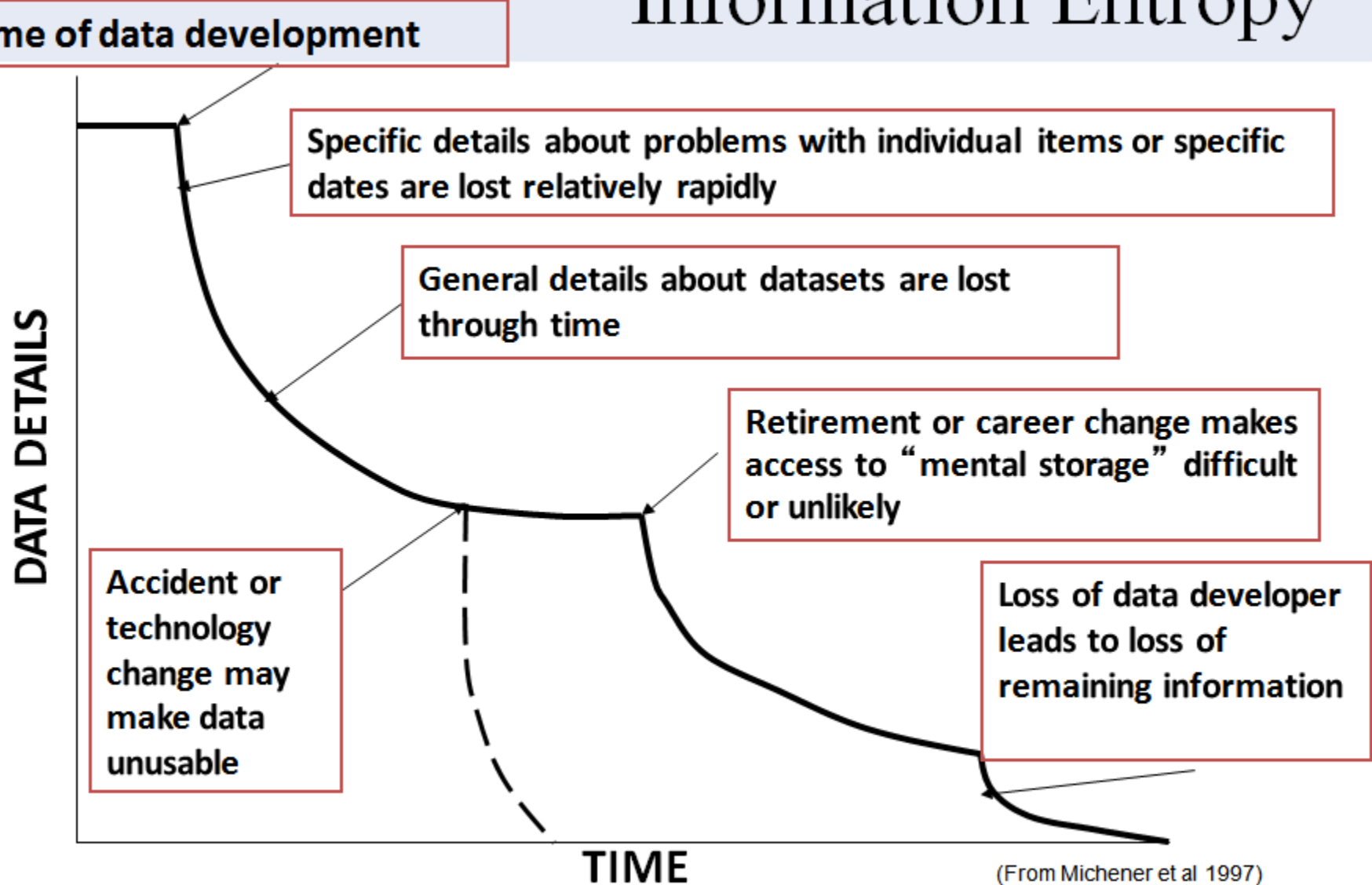
Documentation & Metadata

Agenda:

- Why is documenting your research important?
- Metadata: What is it? Why is it important?
- Creating documentation and metadata:
  - Best practices
  - Tools
- Demonstration using Tools
- Questions

Documentation of Research Data

Discussion

- Look at the few examples of "research" data

- Do you understand what it means?

- Do you know how it was used in the research?

- Is there any documentation?

- What do you need to use this "data"?

Information Entropy

Time of data development

DATA DETAILS

Specific details about problems with individual items or specific dates are lost relatively rapidly

General details about datasets are lost through time

Retirement or career change makes access to "mental storage" difficult or unlikely

Accident or technology change may make data unusable

Loss of data developer leads to loss of remaining information

TIME

(From Michener et al 1997)

Working with Data

- When you *provide* data to someone else, what types of information would you want to include with the data?

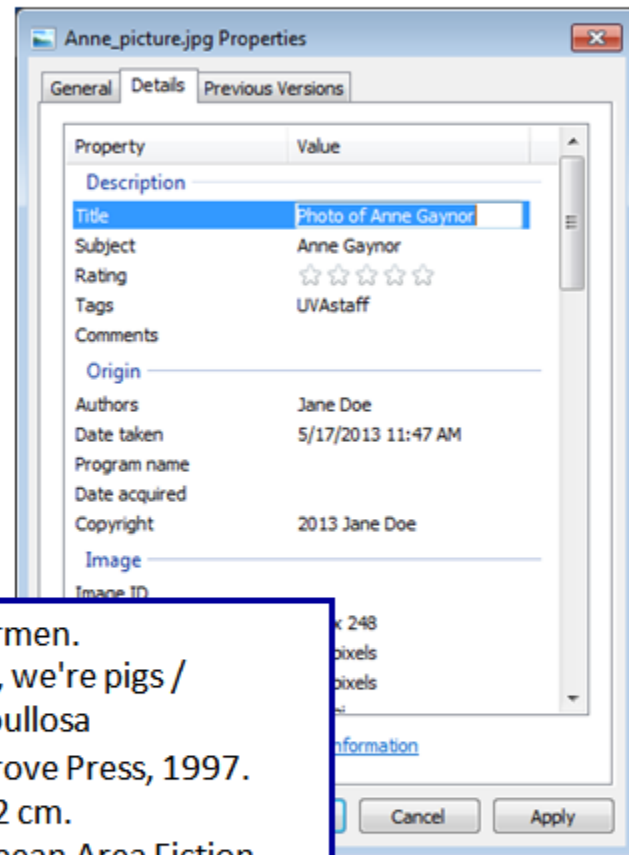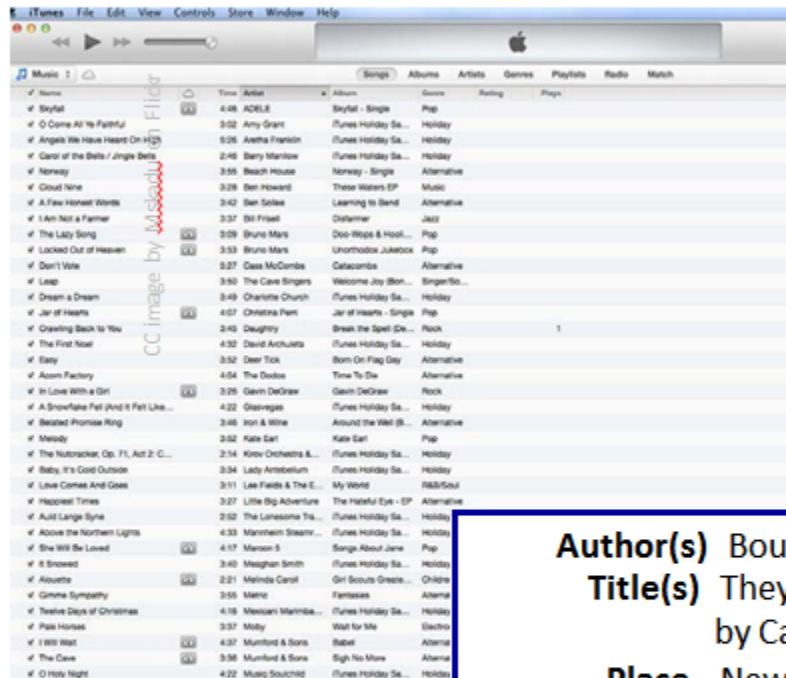- When you *receive* a dataset from an external source, what types of details do you want to know about the data?

Documentation and Metadata Answer…

- **Who** created the data?
- **Who** maintains it?
- **When** were the data collected? **When** were they published?
- **Where** was it collected (geographic location)?
- **What** is the content of the data? The structure?
- **Why** were the data created?
- **How** were they produced/analyzed?

Metadata

- **What is it?**
    - Information that describes a resource

- **Why is it important?**
    - Good metadata will help others understand and use your data
    - Enables a resource or data to be easily discovered

Metadata in Everyday Life

UNIVERSITY _of_ VIRGINIA LIBRARY

Metadata in Research

Project Documentation	Dataset Documentation
• Context of data collection • Data collection methods • Structure, organization of data files • Data sources used • Data validation, quality assurance • Transformations of data from the raw data through analysis • Information on confidentiality, access and use conditions	• Variable names and descriptions • Explanation of codes and schemas used • Algorithms used to transform data • File format and software (including version) used

Critical Roles of Metadata

- **Data Discovery**
  - To be able to identify important data sets
- **Data Retrieval**
  - To know how and where to access data
- **Data Use**
  - To know enough details about how the how the data were collected and stored
- **Data Archiving**
  - Data can grow more valuable with time, but only if the critical information required to retrieve and interpret the data remains available

Metadata Formats

- Documentation for understanding & re-use
  - Readme File
  - Data Dictionary
  - Codebook

- Structured metadata in XML format for use in programs
  - DDI
  - FGDC
  - EML

Unstructured Documentation

- ## Data Dictionary
  http://people.virginia.edu/~sah/bsel/DataDefinitions.pdf

- ## ReadMe File
  http://libra.lib.virginia.edu/dataset_readme_template

- ## Dryad Example (lab notebook)
  http://wiki.datadryad.org/wg/dryad/images/3/3d/DryadLab_example_readme.pdf

Data Dictionary Creation

Exercise

- Choose a dataset
- Use the data dictionary template on the workshop web page
- Add variable names, add descriptions
- What additional information would be needed (units, format information)?
- Is this Data Dictionary enough documentation?

Semi-Structured Documentation

- ## Commented code files
    - .do file (SAS, Stata)
    - .r file
    - .py (Python code)

Structured XML

Standard Schemes (XML)

   – DDI– Data Document Initiative

   http://www.ddialliance.org/

   – FGDC– Geospatial Metadata Standard

   http://www.fgdc.gov/metadata/geospatial-metadata-
   standards

   – EML– Ecological Metadata Language

   http://knb.ecoinformatics.org/software/eml/

Critical Roles of Metadata

- **Data Discovery**
  - To be able to identify important data sets
- **Data Retrieval**
  - To know how and where to access data
- **Data Use**
  - To know enough details about how the how the data were collected and stored
- **Data Archiving**
  - Data can grow more valuable with time, but only if the critical information required to retrieve and interpret the data remains available

Metadata Standards

Standard Vocabulary

- Controlled vocabulary
  - MeSH
  - [DDI Vocabularies](#)
- Standard codes
- Standard formats (date/time/geo-spatial)

  ISO 8601 – YYYY-MM-DDThh:mm:ss.sTZD

  1997-07-16T19:20:30.45+01:00

  Spatial Coordinates for Latitute/Longitude   +/- DD.DDDDD
  -78.476 (longitude)
  +38.029 (latitude)

Structured Metadata Tools

Tools

- Colectica add-on for Excel (DDI)
- Nesstar (DDI)
- Metavist (FGDC)
- ArcGIS (FGDC) *
- Morpho (EML)

http://data.library.virginia.edu/data-management/plan/metadata/metadata-workshop/

Colectica for Excel

- Excel Addin (DDI)

- Describes data files, variables, and code listings (metadata saved in the excel file)

- Import SPSS (.sav) & Stata (.dta) files into Excel, along with metadata

- Code books can also be customized and generated by the tool with various outputs

http://www.colectica.com/software/colecticaforexcel

Nesstar Publisher

- DDI Metadata Editor
- Creates codebooks

http://www.nesstar.com/software/publisher.html

Metavist

- Metadata editor for FGDC

- Includes fields for the Biological Data Profile

http://metavist.djames.net/

ArcGIS

- ArcInfo suite includes ArcCatalog, a tool for organizing GIS data and recording metadata

Morpho Data Management Software

- Creates EML metadata

- Create a catalog of data & metadata upon which to query, edit and view data collection

- Easy-to-use, cross-platform application for local and network access

http://knb.ecoinformatics.org/morphoportal.jsp

QUESTIONS?

Anne Gaynor
Non-English Language Metadata Libraria
Metadata Management Services
amgaynor@virginia.edu

Sherry Lake
Senior Data Consultant
Research Data Services
shLake@virginia.edu

Research Data Services, Data Management Consulting Group
http://dmconsult.library.virginia.edu/training-sessions/