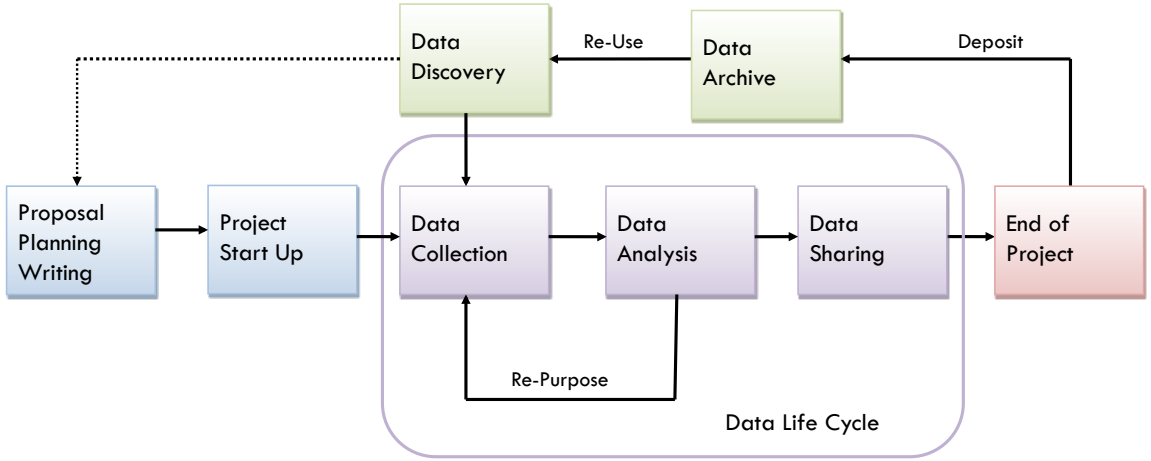
Data Wrangling and Interoperability



Ricky Patterson
Data Management Consulting Group
University of Virginia Library
ricky@virginia.edu

Andrea Denton
Research and Data Services Manager
Claude Moore Health Sciences Library
ash6b@virginia.edu

Goals for the workshop

- Learn about challenges of interoperability
- Understand differences between software that appears to be the same (Excel, Google Spreadsheets, etc.)
- Learn about Open Refine as a tool to fix messy data
- Gain peer and expert feedback

Challenges of interoperability

- Data sets can be isolated, fragmented
- Challenge of combining disparate data sets
  - Formats (proprietary and open)
  - Data definitions (units, time steps, etc.)
  - Missing or poorly formed data

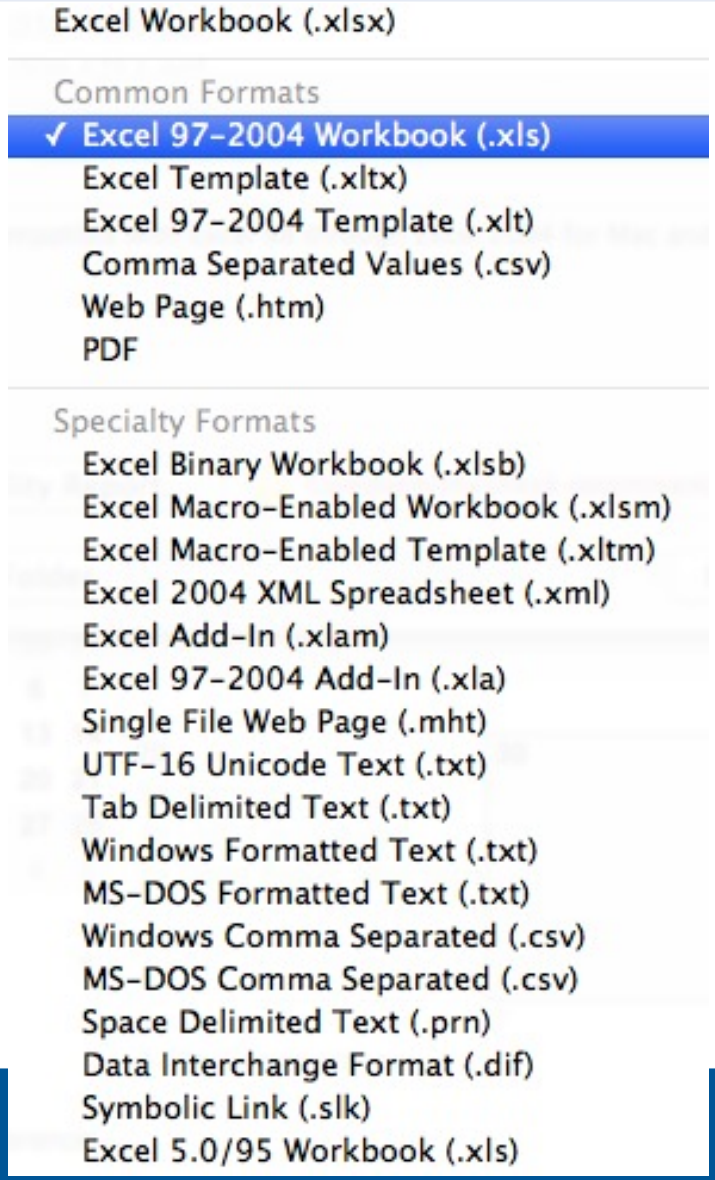
Interoperability with spreadsheets

- Show sample Excel files
- Show sample Google Spreadsheet files
- Demonstrate format save issues
- Show import/export examples and process

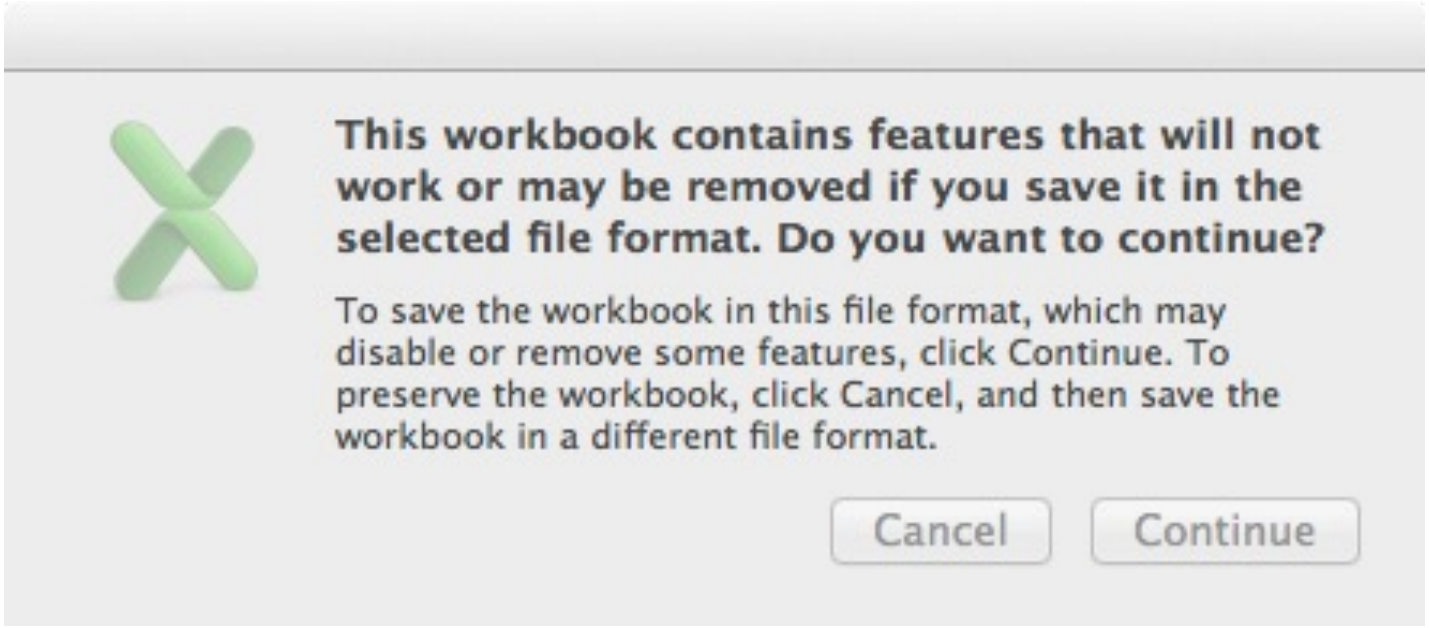
Interoperability with spreadsheets

- Excel
  - Excel 2003 (.xls format)
  - Excel 2007 (.xlsx format begins)
  - **Excel 2010 (Windows)**
  - **Excel 2011 (Macintosh)**
- “Save As” Options

Interoperability with spreadsheets



Interoperability with spreadsheets



Activity 1: Excel import/export problems

- Give some files that have problems
- Ask them to take our csv file, upload it
- Plant some specific traps for them with misaligned fields, nulls, stuff like that
- Take old versions and have them try to upload those
- Have students evaluate export choices and evaluate interoperability

Lessons from Activity 1

- Highlight different points and lessons

Activity 2: Google Spreadsheets import/export problems

- Give some files that have problems
- Ask them to take our csv file, upload it
- Plant some specific traps for them with misaligned fields, nulls, stuff like that
- Take old versions and have them try to upload those
- Have students evaluate export choices and evaluate interoperability

Wrangling messy data

- Cleaning a single data set can be tough, but merging two disparate data sets can be much harder
- Thinking about basic organizational best practices makes all of it easier

Messy issues

- Identification of relationship in data
- Different units across data sets
- Missing data
- Inequivalent time steps, resolution, grid size, dimensions, scale.
- Transformation
- Assumptions, judgment, accuracy from data collection

Open Refine

Google refine

- Formerly Google Refine
- Java applet, web application, but it runs completely locally. Nothing is in the cloud – safe for sensitive data
- No longer Google project,

now just called Open Refine

Open Refine

- Open Refine is used to clean up files, such as spreadsheets, not to create them
  - Take existing excel files, and refine them
- Uses JSON for scripting
  - Clean up one spreadsheet
  - Apply same actions to a different spreadsheet
    - Can select a subset of these actions

Open Refine

- Two exercises:
  - www.freeyourmetadata.org/cleanup
  - http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

Activity 3: use Open Refine

- Load 2 or more messy files (suggest specific locations to find messy files)
- Clean the files
- Merge the files [FIGURE OUT WHERE]
- Export a clean version at the end

Discuss lessons learned

- Open discussion time to talk about issues
- Ask questions
- Talk about other experiences of messy data
- Stress why best practice principles matter

Best Practices Creating Data

1. Use Consistent Data Organization
2. Use Standardized Formats
3. Assign Descriptive File Names
4. Perform Basic Quality Assurance / Quality Control
5. Preserve Information - Use Scripted Languages
6. Define Contents of Data Files; Create Metadata
7. Use Consistent, Stable and Open File Formats

Mailing List Subscription

- Please check the box on our sign-in sheet to receive occasional emails to keep up with our services, training, and news.

- Please encourage others to subscribe: http://eepurl.com/CJwYT

More Research Data Services in the Library

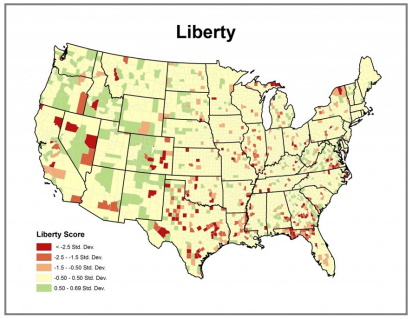
Offering expert data assistance at every stage of the research process.

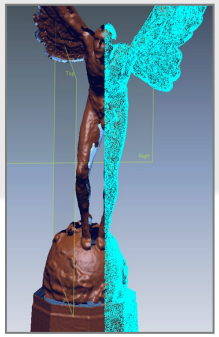
PLANNING

Need a data management plan?

We can assist you with developing a data management plan that meets increasingly stringent criteria from funding agencies, including:

- Implementation of procedures, tools, and workflows for managing data sets
- Designing a strong study that yields reliable statistics





FINDING & COLLECTING

Need help finding data or collecting your own?

We have thousands of sources with the data you seek and experts who will help you:

- Locate, evaluate and format data
- Design metadata and data documentation protocols for new data collection
- Capture data with the appropriate technology tools for your needs

SHARING

Ready to share or archive your data?

We can consult with you on strategies to help others discover or access your research by:

- Adhering to data sharing policies and norms
- Selecting a data-sharing repository
- Making your data easier to discover and link

ANALYZING

Want help uncovering unique and compelling insights?

Get expert assistance from statistical, spatial, or media specialists to analyze your data and convey your research message:

- Learn how to use cutting-edge tools and methods
- Experiment with high-resolution visualization technologies
- Develop graphical representations that bring impact to your analysis

University of Virginia Library

Workshops • 1:1 Consultations • Class Presentations

Contact me at als9q@virginia.edu to find out more.

QUESTIONS?

Ricky Patterson
Data Management Consulting Group
University of Virginia Library
ricky@virginia.edu

Data Management Consulting Group
University of Virginia Library
http://dmconsult.library.virginia.edu