Dec 11th, 1:30 PM

Check and Understand Presentation

Heidi Imker
University of Illinois, Urbana-Champaign, imker@illinois.edu

C ⇛ **U** ⇛ R ⇛ A ⇛ T ⇛ E

⇓ ⇓

Check & Understand

So … in comes a dataset … now what?

Check Step - Action

- Review the content of the data files (e.g., open and run the files or code).
- Verify all metadata provided by the author and review the available documentation.

Check Step - Curator Checklist

- ❑ Files open as expected
  - ❑ Issues _____
- ❑ Code runs as expected
  - ❑ Produces minor errors
  - ❑ Does not run and/or produces many errors
- ❑ Metadata quality is rich, accurate, and complete
  - ❑ Metadata has issues _____
- ❑ Documentation Type (circle)

  Readme / Codebook / Data Dictionary / Other: _____
  - ❑ Missing/None
  - ❑ Needs work

Understand Step - Action

- Check for quality assurance and usability issues such as missing data, ambiguous headings, code execution failures, and data presentation concerns.
- Try to detect and extract any “hidden documentation” inherent to the data files that may facilitate reuse.
- Determine if the documentation of the data is sufficient for a user with similar qualifications to the author’s to understand and reuse the data. If not, recommend or create additional documentation (e.g., a readme.txt template).

Understand Step - Curator Checklist

Varies based on file formats and subject domain. For example….

Tabular Data Questions (Microsoft Excel)

- ❑ Organization of data well-structured
  - ❑ Not rectangular
  - ❑ Split tables into separate tabs
- ❑ Headers/codes clearly defined
  - ❑ Define headers
  - ❑ Clarify codes used _____
  - ❑ Clarify use of “blanks”
  - ❑ Clarify units of measurement
- ❑ Quality control clearly defined
  - ❑ Unclear quality control
  - ❑ Update/add Methodology

Start!

Your datasets are here: http://bit.ly/2klXVyq

Wait!! I need more structure…

DeepBlueData_QualtricsReviewForm_20171017.PDF

https://drive.google.com/open?id=0By0h3bM2cP5wUDZLWUxhM0p5Y3M

Documentation can be for…

- Explaining and/or maintaining consistency of data
- Training new people
- Assessing data for reuse
- Assistance in actual reuse
- Efficiency in archiving

Data Documentation Continuum

README Template

```
README TEMPLATE

Dataset Title:

Author(s):

Description:

Methods and Processing:

Funding Source:

Supplements:
```

```
<?xml version="1.0"?>

<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <dc:title>
    UKOLN
  </dc:title>
  <dc:description>
    UKOLN is a national focus of expertise in digital information
    management. It provides policy, research and awareness services
    to the UK library, information and cultural heritage communities.
    UKOLN is based at the University of Bath.
  </dc:description>
  <dc:publisher>
    UKOLN, University of Bath
  </dc:publisher>
  <dc:identifier>
    http://www.ukoln.ac.uk/
  </dc:identifier>

</metadata>
```

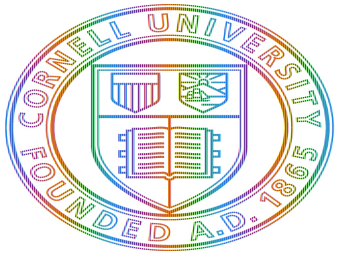
Note that the http://example.org/myapp/schema.xsd XML schema does not exist - this is a ficticious example.

Informal ReadMe → **Formal Schema**

Lower-Barrier	Lower-Potential	Higher-Potential	Higher-Barrier
Fast	Irregular	Standardized	Slow
Easy	Incomplete	Machine actionable	Skilled

ReadMe Files

- Swiss-army knife of documentation
- Semi-structured, human-readable text files
- Provide basic information, such as:
  - author, year, associated publications
  - accounts for all files and folders in a dataset
  - explanation of naming conventions
  - relationship between directory structure and the data
  - licensing and/or reuse information



Cornell Un
Library

Guide to writing "readme" style metadata

- Recommended content
  - General information
  - Data and file overview
  - Sharing and access information
  - Methodological information
  - Data-specific information

https://data.research.cornell.edu/content/readme

Unavoidable Challenges

Some Issues with Check

- Can’t open the files/code
- Too many files to open
- Can open the files, but don’t know what I’m looking for
- No (or not enough) documentation

Some Issues with Understanding

- Not enough experience to evaluate these data
- Documentation is over my head
- Still can’t tell what’s typical in this field
- Required software is too hard to find/use/justify $$$