

Jul 12th, 10:45 AM - 11:45 AM

Introducing Undergraduates to Research Datasets

Bill G. Kelm

Willamette University, bkelm@willamette.edu

John Repplinger

Willamette University

Let us know how access to this document benefits you.

Follow this and additional works at: <http://pdxscholar.library.pdx.edu/nwirug>

Bill G. Kelm and John Repplinger, "Introducing Undergraduates to Research Datasets" (July 12, 2016). *Northwest IR User Group*. Paper 4.

<http://pdxscholar.library.pdx.edu/nwirug/2016/Presentations/4>

This Panel Discussion is brought to you for free and open access. It has been accepted for inclusion in Northwest IR User Group by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.

Introducing Undergraduates to Research Datasets

Bill Kelm & John Reppinger

Willamette University

What We're Covering...

Bill Kelm (Behind Scenes)

- Initial Concept and Discussions
- Metadata Elements
- The Experts
- Students Documenting Data
- Delivery Options
- Sample File
- Customize DSpace

John Replinger (Educating)

- Coordinate with Instructor
- Meet the Students
- End-Semester Scramble
- Specific Issues
- Success and Failure
- The Aftermath
- Let's Improve...

Initial Dataset Discussions

- NW5C Summer Workshop
 - June 2015 at Lewis & Clark:
faculty / librarians / students
- Return with a plan for Spring 2016



(Image source: reed.edu)

Metadata Elements in README

- DMPTool

https://dmptool.org/dm_guidance#metadata

- Best Practices Cornell

<http://data.research.cornell.edu/content/readme#bestpractices>

- Discussions

- Do we need elements already captured by DSpace?
- Are there some elements we will not have?
- Do we then even need a readme file if we keep data with theses?



(Image source: www.flickr.com/photos/comedynose)

Local / National Experts

- Steve Tuyl (Oregon State)
 - “the metadata associated with the DSpace repository item record is there for discovery purposes and administrative purposes, while the metadata that might “ride alongside” the dataset itself is there for usability purposes.”
- Research Data Management and Sharing (COURSEA)
<https://www.coursera.org/learn/data-management>



(Image source: www.flickr.com/photos/ian_munroe)

Students Documenting Data

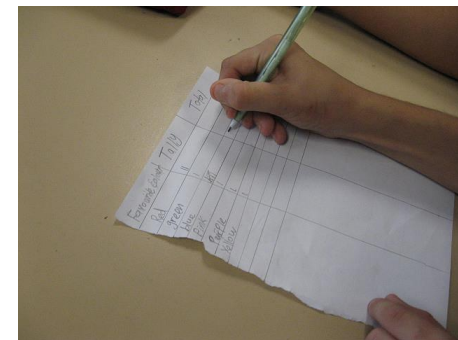
- Some elements would apply and others would not.

General Overview

Title	Name of the dataset or research project that produced it
Creator	Names and addresses of the organizations or people who created the data; preferred format for personal names is surname first (e.g., Smith, Jane).

Technical Description

File inventory	All files associated with the project, including extensions (e.g. 'NWPalaceTR.WRL', 'stone.mov')
File Formats	Formats of the data, e.g., FITS, SPSS, HTML, JPEG, etc.
File structure	Organization of the data file(s) and layout of the variables, where applicable
Version	Unique date/time stamp and identifier for each version
Checksum	A digest value computed for each file that can be used to detect changes; if a recomputed digest differs from the stored digest, the file must have changed



(Image source: www.flickr.com/photos/47893483@N06)

Delivery Options

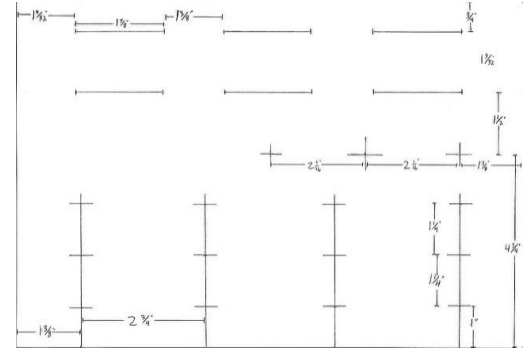
- Needed the best way to present a template
- Found Georgia Tech
 - <http://d7.library.gatech.edu/research-data/readme>
- Editable Word Template vs. Google Doc
- Guide on a Side
 - http://library.willamette.edu/guide_on_the_side/tutorial/academic-commons-data-submission



(Image source: www.flickr.com/photos/vagabondblogger/)

Sample readme.txt File

- Blank Template Issues
- Cornell Sample File



(Image source: <https://www.flickr.com/photos/systemf/2919321523>)

- http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf

- Edited and Brought to Undergrad Level

DSpace Customizations

- Workflow
 - Streamlined questions for just datasets using XMLUI

Describe Item

Authors:

Enter the names of the authors of this item below.

Last First

name, e.g. *Smith* name(s) + "Jr", e.g. *Donald Jr*

Title:

Enter the main title of the data.

Date of Submission:

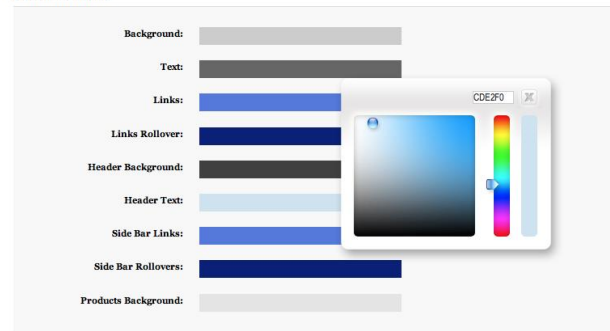
Please give the date that you are submitting your dataset to the Academic Commons.

Year Month Day

Type:

Select the type of content of the item.

Customize Theme



(Image source: www.flickr.com/photos/factoryjoe/370483340)

DSpace Customizations

- Need to provide link between theses and dataset
 - dc.relation.haspart for a link from thesis to the dataset
 - dc.relation.ispartof for a link from dataset to the thesis



(Image source: www.flickr.com/photos/volvob12b/9519733893)

Employing BAC-reporter constructs in the sea anemone *Nematostella vectensis*



Citable URI

<http://hdl.handle.net/1912/6199>

As published

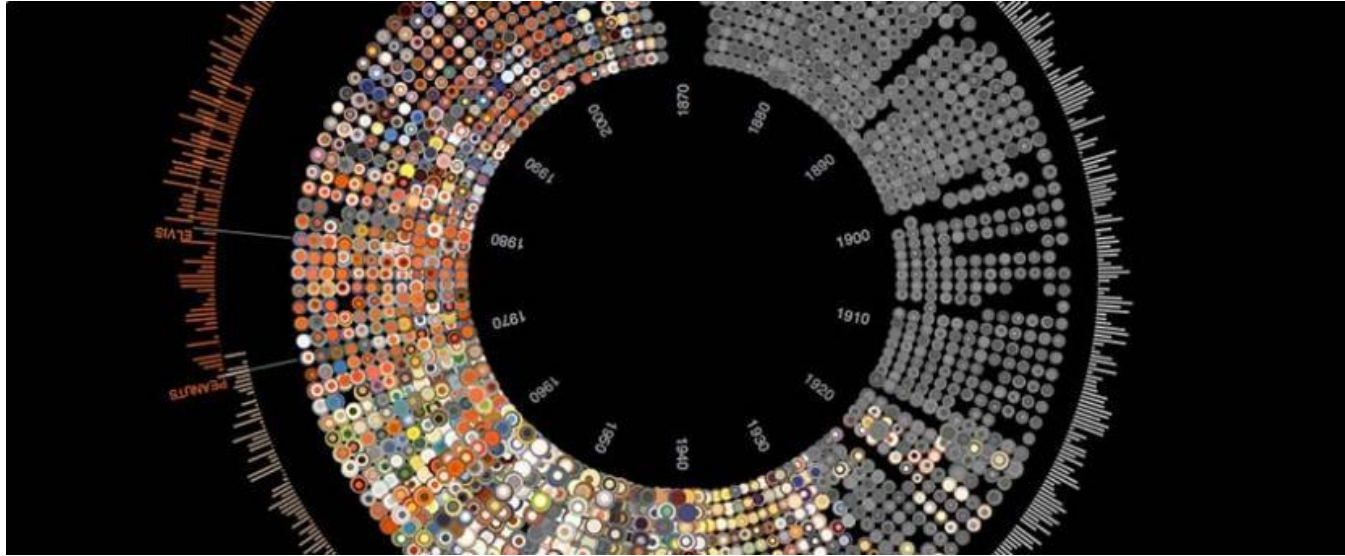
<http://dx.doi.org/10.1093/icb/ict091>

Related Material/Data

<http://hdl.handle.net/1912/6068>

(Image source: <http://darchive.mblwhoilibrary.org/handle/1912/6199>)

Educating... (John)



(Image source: libmedia.willamette.edu)

Coordinate with Instructor

- Librarians met with EES instructor (Jan.)
- All seniors required a dataset w/ thesis
- Instructor provided 1-2 student datasets to play with
- Instructor used different “data management plan” standards
Only 4 criteria (review in class mid-semester):
 - Collection / acquisition of data
 - Data type description
 - Instrument / collection approach
 - Processing & analysis data



(Image source: libmedia.willamette.edu)

Meet the Students

- I met twice with 19 EES seniors (Sep. & Apr. 26)
 - Importance of data management
 - Walk through Readme & file uploads
 - Handout and Guide on the Side
- Library-supplied criteria (Apr.)
 - 13 fields captured
 - Simplified Data Management Plan for faculty
- Student interest in process
 - Confusion with CSV files
 - Not put into practice what they learned



(Image source: sunjournal.com)

End-Semester Scrabble

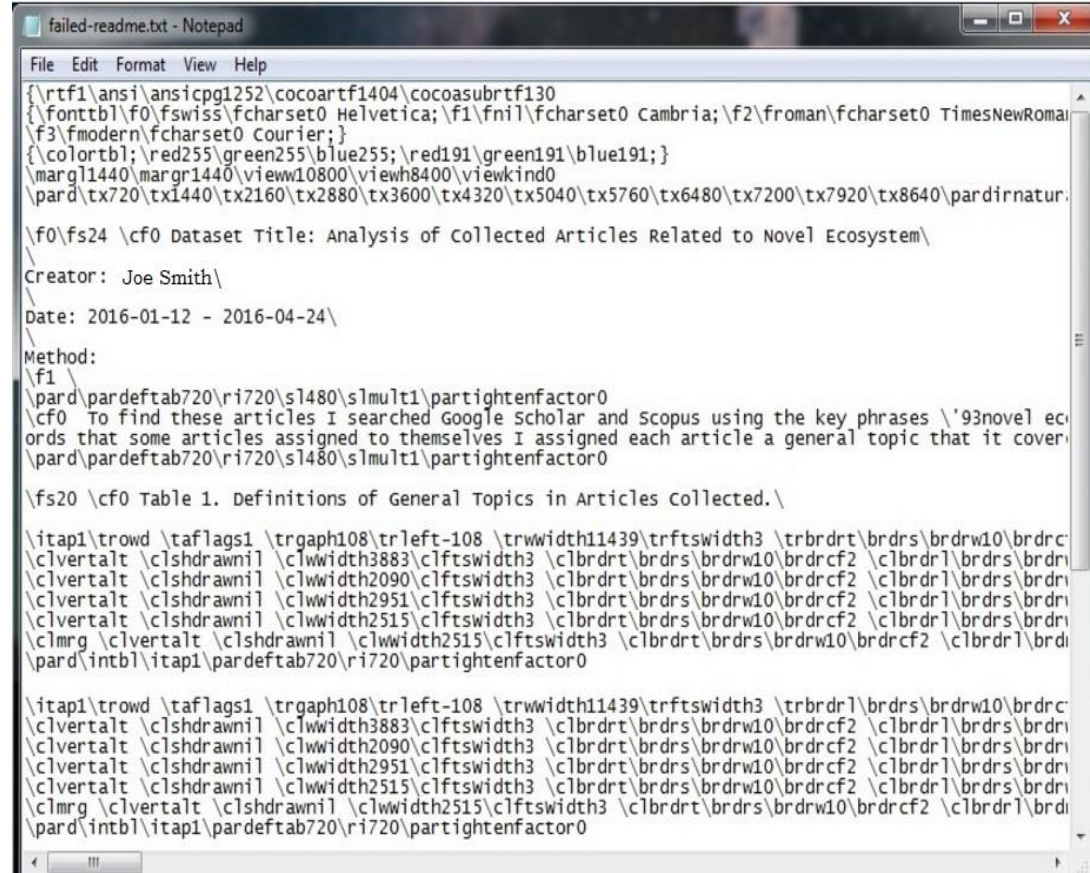
- Typical procrastination
- First few submissions okay...
- Claimed to understand, but really didn't
($\frac{1}{3}$ submissions needed major correction. $\frac{1}{3}$ were great!)
- Some followup appointments & emails
(~ 6-7 related questions)
- Not lack of help, but lack of effort (“not graded, so...”)



(Image source: www.theodyessyonline.com)

Specific Issues

- *Multiple submissions for multiple files
- Lack of metadata
- Subject terms were too few/broad
- Incorrect source info / file structure
- External formatting (copy & paste)
- No info about processing
- No variable list or codes (or worse... partial!!!)



```
failed-readme.txt - Notepad
File Edit Format View Help
{\rtf1\ansi\ansicpg1252\cocoartf1404\cocoasubrtf130
\fonttbl\f0\fswiss\fcharset0 Helvetica;\f1\fnil\fcharset0 Cambria;\f2\froman\fcharset0 TimesNewRoman
\f3\modern\fcharset0 Courier;}
{\colortbl;\red255\green255\blue255;\red191\green191\blue191;}
\margl1440\margr1440\vieww10800\viewh8400\viewkind0
\pard\tx720\txi440\tx2160\tx2880\tx3600\tx4320\tx5040\tx5760\tx6480\tx7200\tx7920\tx8640\pardirnatur
\F0\fs24 \cf0 Dataset Title: Analysis of Collected Articles Related to Novel Ecosystem\
Creator: Joe Smith\
Date: 2016-01-12 - 2016-04-24\
Method:
\fi
\pard\pardefstab720\ri720\sl480\smult1\partightenfactor0
\cf0 To find these articles I searched Google Scholar and Scopus using the key phrases '\93novel eco
ords that some articles assigned to themselves I assigned each article a general topic that it cover
\pard\pardefstab720\ri720\sl480\smult1\partightenfactor0
\fs20 \cf0 Table 1. Definitions of General Topics in Articles collected.\
\itap1\trowd \taflags1 \trgaph108\trleft-108 \trwidth11439\trfstwidth3 \trbrdr\brdrs\brdrw10\brdr
\clvertalt \clshdrawnil \clwidth3883\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clvertalt \clshdrawnil \clwidth2090\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clvertalt \clshdrawnil \clwidth2951\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clvertalt \clshdrawnil \clwidth2515\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clmrg \clvertalt \clshdrawnil \clwidth2515\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdr
\pard\intbl\itap1\pardefstab720\ri720\partightenfactor0
\itap1\trowd \taflags1 \trgaph108\trleft-108 \trwidth11439\trfstwidth3 \trbrdr1\brdrs\brdrw10\brdr
\clvertalt \clshdrawnil \clwidth3883\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clvertalt \clshdrawnil \clwidth2090\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clvertalt \clshdrawnil \clwidth2951\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clvertalt \clshdrawnil \clwidth2515\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdrs\brdr
\clmrg \clvertalt \clshdrawnil \clwidth2515\clfstwidth3 \clbrdr\brdrs\brdrw10\brdr cf2 \clbrdr1\brdr
\pard\intbl\itap1\pardefstab720\ri720\partightenfactor0
```


Success!

“Good” README File
Example

```
Untitled - Notepad
File Edit Format View Help
General Overview
Dataset Title: Alkalinity Titrations for Ponil Creek Watershed New Mexico
Creator: Hansen, Cassandra
Date: 2016-05-02
Method: I collected 54 water samples from the three branches of the Ponil Creek and prefo
GS, 2013). These alkalinity titrations will provide the chemical concentration for the w
Processing: Data was manually entered from field notebook into an Excel spreadsheet and e
Source: No external data sources are referenced.
Funder: No funding was provided. Equipment provided by Katja Meyer of Willamette EES Depa
Thesis Title: controls on weathering in the Ponil Creek watershed, New Mexico

Content Description
Subject: Alkalinity titrations and chemical weathering fluxes for Ponil Creek Watershed N
Place: Data was collected in Cimarron, New Mexico in the Ponil Creek watershed. See UTM c
Variable List and Codes:

Variables
Sample ID: simplistic way to refer to specific samples, numbered in order of collection.
River: Refers to the branch of the Ponil Creek of the study site. There are three branche
Location: A common name for the location if it has one, otherwise made up name for sampl
Date: Given in short hand the first digit being the month (6=June, 7=July, 8=August) and
Time taken: All times are in MDT (Mountain Daylight Time) UTC/GMT -6 hours and were recor
UTM: Given in UTM coordinates using a Delorme Earthmate PN-40 GPS set to UTM/UPS Coordin
Temperature: Temperature of the Creek at time of sampling taken by using a Thermo Scient

Specific Conductance: Conductivity of the creek at time of sampling taken by using a Ther
Time analyzed: Time that the alkalinity titration was performed later on the same day as
Sulfuric Acid concentration: The sulfuric acid used was a HACH Company Digital Titration
Titration Data: Titrations were completed the same day as the sample was taken. A magnet
pH: pH was measured using a Thermo Scientific Orion Star A325 pH/Conductivity Meter with
Clicks: clicks of sulfuric acid delivered during the course of the titration, the HACH D
mg/L of CaCO3: using the USGS web Based Alkalinity calculator version 2.22 oregon water

Photo: Taken by an iPhone 5 at the sample site at the time of sample collection.
Technical Description
File structure: CJH Thesis Field Research.xlsx - The original Microsoft Excel spreadsheet

File #1 * AllData.cvs;
File #2 - weatheringFlux.cvs;
File #3 - LatLong;
File #4 * PaperFluxes.cvs;
File #5 * upstreamdownstream.cvs;
File #6 * Flood.cvs;
File #7 * Fire.cvs;
File #8 * USGSFlood.cvs

Necessary software: Excel 98 or more recent is needed to view and access the CJH Thesis F
Access Rights: This data is freely available for re-use. Please acknowledge Cassandra Har
Other Notes: No other notes are needed.
```

The Aftermath

- Overall process went as expected
- Weeks to makes corrections (~ two weeks)
- Link theses and datasets (~ two day)
Thesis used: `dc.relation.haspart` + URL to link to the dataset
Datasets used: `dc.relation.ispartof` + URL to link to the thesis



(Image source: <https://i.ytimg.com/vi/dH5zkYjgnKA/maxresdefault.jpg>)

Let's Improve...

- Instructors & Library must use same data standards
- Introduce students to standards BEFORE collecting data
- Instructors work with students to fill out metadata throughout research
- Library to provide good/bad README file metadata examples
(separate DSpace collection & for all disciplines)
- Library to provide tool for CSV files
(separate tabs within Excel and zip files together again)
- Decide core competencies for data management
(Suggestions welcome)

Questions & Comments?

Bill Kelm bkelm@willamette.edu

John Replinger jreplin@willamette.edu

README File Criteria

<http://libmedia.willamette.edu/info/data/student.html> (Google Doc)

General Overview

Dataset Title:

Creator:

Date:

Method:

Processing:

Source:

Funder:

Thesis Title:

Content Description

Subject:

Place:

Variable List and Codes:

Technical Description

File structure:

Necessary software:

Access Rights

Rights:

Other Notes

Notes:

Thesis Advisor:

Note: Enter "N/A" for fields that do not apply to your project.