BEYOND METADATA: LEVERAGING THE "README" TO SUPPORT DISCIPLINARY DOCUMENTATION NEEDS

Lizzy Rolando
Georgia Tech Library
RDAP 2015

WHAT WE DO NOW (PT. 1)

- Data archiving in Institutional Repository, SMARTech
- Dublin Core for metadata record

Cognitive Process Model, Validation Data, Initial Modeling Results

Show simple item record

dc.contributor.advisor	Feigh, Karen M.
dc.contributor.author	Chua, Zarrin K.
dc.date.accessioned	2013-07-30T19:26:31Z
dc.date.available	2014-07-31T05:30:05Z
dc.date.issued	2013-07-30
dc.identifier.uri	http://hdl.handle.net/1853/48588
dc.description	The files associated with this project are separated into two groups. The first ("EvaluationLunarLandingAutomation_ExperimentalFiles.zip") contains all of the files used to design an experiment conducted at Johnson Space Center with the NASA Astronaut Office in August 2012. The area of investigation for this experiment was Astronaut decision-making processes during lunar landing in areas around the South Pole of the Moon. The files are packaged in a zipped file, and they must be extracted before they can be used.
dc.description	The second set of files ("CognitiveProcessModel.zip") contains all of the files used to create the cognitive process model (moderate, Apollo-like function allocations) and four landing areas on the South Pole of the moon, for thesis, validation, and initial results. The model uses data from four landing areas. The period of development was January 1, 2013 - March 1, 2013, and the area of investigation was the South Pole of the Moon. With this data set, the user should be able to visualize the chosen landing sites for each user in the August 2012 human in the loop experiment conducted with the NASA astronaut office, validation of the cognitive model, and a set of randomly generated data points used for initial results. These data and models are discussed in the dissertation "System design considerations for human-automation function allocation during lunar landing," which will be available in SMARTech in the coming months. The files comprising the model are packaged in a zipped file, and they must be extracted before they can be used. MATLAB software is required to run the model.
dc.description.abstract	These are the model files for the cognitive process model (moderate, Apollo-like function allocations) and four landing areas on the South Pole of the moon. With this data set, the user should be able to visualize the chosen landing sites for each user in the August 2012 human in the loop experiment conducted with the NASA astronaut office, validation of the cognitive model, and a set of randomly generated data points used for initial results.
dc.description.sponsorship	United States. National Aeronautics and Space Administration
dc.language.iso	en_US
dc.publisher	Georgia Institute of Technology
dc.subject	Cognitive model
dc.subject	Astronaut decision making
dc.subject	Landing point designation
dc.subject	Lunar landing
dc.title	Cognitive Process Model, Validation Data, Initial Modeling Results
dc.type	Dataset
dc.contributor.corporatename	Georgia Institute of Technology. School of Aerospace Engineering
dc.contributor.corporatename	Georgia Institute of Technology. Cognitive Engineering Center
dc.embargo.terms	1 year

WHAT WE DO NOW (PT. 2)

- Require “README.txt” to capture information not well suited for SMARTech records
- Supply template to depositors to help with creating README

```
Genotype and fertility raw data from:

Brown KM, Burk LM, Henagan LM, Noor MAF (2004). A test of the chromosomal rearrangement model of speciation in
Drosophila pseudoobscura. Evolution 58(8): 1856-1860.

Six data files are included- one for each backcross.  The crosses performed were between Drosophila persimilis
strain MSH 1993 and either a D. pseudoobscura bogotana or a D. pseudoobscura pseudoobscura strain (see Brown
et al 2004).  As described in the paper, the cross was performed as:
persimilis male x (ps or bog) female -> F1 female x (ps or bog) male

Data files are comma-delimited, and each file has 6 columns:

1) DissectionDate (8 days after eclosion from pupal case)
2) Number (order of dissection on dissection date)
3) Whether the fly was fertile (F) or sterile (S)
4) Genotype of marker on XL chromosome arm
5) Genotype of marker on XR chromosome arm
6) Genotype of marker on second chromosome

The last three column headers also indicate the marker or markers used for genotyping in that backcross.

Genotypes are listed as "b" (for "bog"), "per", or "ps."  For the X-chromosome markers, the backcross males
were hemizygous, so it indicates the actual hemizygous genotype.  For the 2-chromosome markers, the backcross
males were homozygous or heterozygous.  In that case, a genotype of "per" indicates a heterozygote, whereas
genotypes of "b" or "ps" are homozygous.

The data file for the white-eye mutant subculture of the el Recreo (ER) line is supplemented with extensive
data collected after publication of the Brown et al (2004) study, as used in Chang and Noor (2007).  The
result relative to Brown et al (2004) is unchanged.

Also relevant to the ER line- the flies used were selected for bearing the white mutation (on the XL
chromosome arm).  Hence, far fewer flies bore the D. persimilis allele at the XL chromosome arm marker.

The data from the Flagstaff 1993 line were recycled from an earlier study (Noor et al 2001).
```

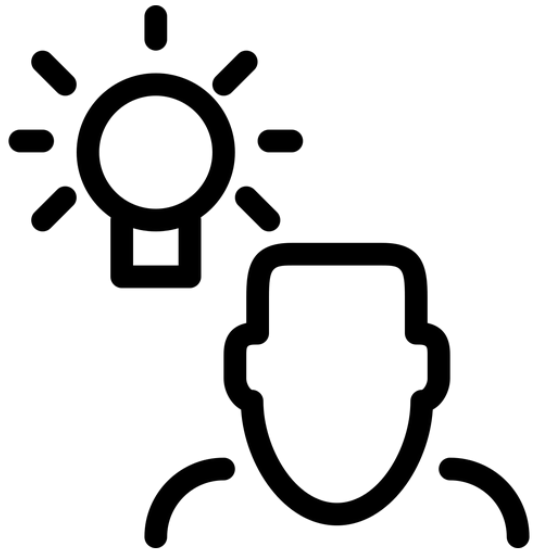
Example README from (Brown KM, Burk LM, Henagan LM, Noor MAF, 2004)

WHAT WASN'T WORKING

Image from Xerox ad: https://editbarry.files.wordpress.com/2011/07/12_xerox_s1.jpeg

PROPOSED SOLUTION

- Disciplinary templates
  - Work with subject librarians to engage community
  - Have more specialized information for depositors
  - Accommodate disciplinary needs, even when using a generic, one-size-fits-all repository



Created by iconsmind.com, from Noun Project

METHODS (PT. 1)

Interviews

- 3 Civil & Environmental Engineering
- 3 Interactive Computing
- 2 Economics

Example Questions

- What sorts of information do you record about your data?
- What would someone else need to know about your data in order to use it themselves?
- If you have used someone else's data in the past, what sorts of information did you need in order to evaluate, understand, and reuse those data?

METHODS (PT. 2)

Mine Investigator publications for contextual information

(Chao, 2015)

Geographic Information

Methods and Sampling Information

and dissemination of antibiotic resistance in complex microbial communities. Understanding the interactions and roles of different microbial species in a community under the environmental stress brought about by antimicrobial compounds such as QACs is vital. The objective of the work presented here was to systematically assess the effect of benzalkonium chlorides (BACs), a widely used class of QACs, on the structure of three aerobic microbial communities as well as their antimicrobial resistance.

MATERIALS AND METHODS

Microbial Community Development. Three aerobic microbial communities were developed and maintained in the laboratory for over 4 years with different carbon and energy sources and/or a BAC mixture: (a) dextrin/peptone 50:50 (w/w) mixture (DP); (b) dextrin/peptone plus BAC mixture (DPB); and (c) BAC mixture only (B). The DP microbial community was developed from an inoculum of a contaminated sediment sample collected at the Bayou d'Inde, a tributary of the Calcasieu River, near Lake Charles, LA. After one year of maintaining the DP community, the DPB community was developed with inoculum from the DP community, which in turn was subsequently used as inoculum to develop the B community. The B community was enriched and has been maintained for over 4 years with the BAC mixture as the sole carbon/energy source, supplemented with NH_4NO_3 as the nitrogen source. The BAC mixture consisted of 60:40 (w/w) dodecyl benzyl dimethyl ammonium chloride (C_{12}BDMA-Cl; $C_{21}H_{38}NCl$) and tetradecyl benzyl dimethyl ammonium chloride (C_{14}BDMA-Cl; $C_{23}H_{42}NCl$) purchased from Sigma Aldrich (St. Louis, MO). The three communities were maintained at room temperature (22–24 °C) in aerated, fed-batch reactors with a residence time of 14 days. One-fourth of the culture volume was wasted and replenished twice a week with autoclaved mineral medium along with dextrin/peptone (DP and DPB communities) and/or BAC mixture (DPB and B communities). The medium contained the following (in g/L): K_2HPO_4, 0.6; KH_2PO_4, 0.34; $CaCl_2{\cdot}2H_2O$, 0.07; $MgCl_2{\cdot}6H_2O$, 0.14; $MgSO_4{\cdot}7H_2O$, 0.27; $FeCl_2{\cdot}4H_2O$, 0.07 and 0.7 mL of trace metal stock solution.[13] At the beginning of each feeding cycle, the initial dextrin/peptone concentration was 2200 mg/L, expressed as chemical oxygen demand (COD), and the BAC mixture concentration was 50 mg/L (equivalent to 140 mg COD/L). Upon feeding, the initial NH_4NO_3 concentration in the B community was 44 mg/L. The steady-state pH and volatile suspended solids (VSS) concentration of the DP, DPB, and B communities were 7.7/2500 mg/L, 8.1/1000 mg/L, and 6.9/138 mg/L, respectively. To characterize the feeding cycle of each community, liquid samples were taken at the start of each cycle and at appropriate time intervals to measure pH, soluble COD, and BAC concentration.

Microbial Community Analysis. A clone library-based molecular phylogenetic approach was used to decipher the bacterial community structure in the DP, DPB, and B communities. Samples were collected from each community in the middle of the feeding cycle, washed several times with a saline phosphate buffer, and then genomic DNA was extracted using the Microbial DNA isolation Kit (MO BIO Laboratories, Carlsbad, CA) according to the manufacturer's instructions. The 16S rRNA gene was amplified by PCR using the universal oligonucleotide primer pairs 27F (5′-AGAGTTT-GATCCTGGCTCAG-3′) and 1541R (5′-AAGGAGGT-GATCCARCCGCA-3′). The PCR amplification was carried out with the following conditions: single cycle denaturation at 94 °C for 5 min; 30 cycles of 30 s at 94 °C, 30 s at 55 °C, 2 min at 72 °C, and final extension for 7 min at 72 °C. Cloning was performed using TOPO TA cloning kit (Invitrogen, Carlsbad, CA) as instructed by the manufacturer. The obtained clones were then grouped into different operational taxonomical units (OTUs) after being digested with restriction enzymes *Msp1*/*Taq1* based on the restriction patterns. Representative clones of each OTU were selected for the sequencing of nucleotides after PCR amplification and purification. The obtained 16S rRNA gene sequences were then queried against the NCBI database[14] using the MEGABLAST algorithm. The partial sequences were between 600 and 950 base-pair long. All 16S rRNA gene sequences were manually trimmed of the vector and primer sequence, and the alignment was performed with the program CLUSTALW. The sequence-based phylogenetic tree of the dominant bacteria was constructed by applying the neighbor-joining algorithm with Jukes Cantor correction using the program MEGA5. The tree topology was evaluated by bootstrap resampling analysis of 1000 resampling data sets. To define the community richness and diversity, Chao 1 richness index and Shannon–Wiener diversity index were calculated using FastGroup II.[15] Thirty-eight 16S rRNA gene nucleotide sequences (18 from DP, 11 from DPB, and 9 from B community) have been deposited in the National Center for Biotechnology Information (NCBI-GenBank) under accession numbers KC491136 to KC491173.

Assessment of Antimicrobial Susceptibility. The antimicrobial susceptibility and resistance of the three microbial communities were measured for the BAC mixture, penicillin G, tetracycline, and ciprofloxacin using a macrobroth dilution procedure adapted from the National Committee for Clinical Laboratory Standards using the Mueller-Hinton broth as the carbon and energy source.[16] Inocula were prepared by removing samples from each community and diluting in the Mueller-Hinton broth to an optical density of 0.1 (equivalent to 0.5 McFarland turbidity standard). Test dilution series were prepared using the broth and the test compounds at an initial concentration range from 0 to 500 μg/mL. Control series were identical to the test series without inoculum. All control and test series were prepared in triplicate 13 × 100 mm glass tubes, and then incubated overnight at room temperature (22–23 °C) while mixed using an orbital shaker at 190 rpm. After incubation, the optical density at 600 nm was measured using a HP 8453 UV–visible spectrophotometer (Hewlett-Packard, Palo Alto, CA) with the control tubes as blanks. The minimum inhibitory concentration (MIC) was recorded as the lowest antimicrobial concentration that prevented any measurable growth. The 90% and 50% inhibitory concentrations (IC_{90} and IC_{50}, respectively) were also recorded as the concentrations in which 90% and 50% growth inhibition was observed, respectively. At the end of the incubation, the concentration of BACs and the three antibiotics in all series was determined.

Contribution of Efflux Pump(s) to Antimicrobial Resistance. Thioridazine was used as an efflux pump inhibitor (EPI) in this study. Although thioridazine has been previously used to study the role of efflux pumps with a diverse group of both antimicrobials and bacteria,[17−19] thioridazine is also toxic. Therefore, we first determined the inhibitory effect of thioridazine to DP, DPB, and B microbial communities at an initial concentration range from 0 to 500 μg/mL using the above-described macrobroth dilution method. On the basis of this preliminary assay, a thioridazine concentration of 30 μg/

METHODS (PT. 3)

Review existing disciplinary metadata standards

Metadata Specification	Disciplinary Coverage
Data Documentation Initiative (DDI)	Social Sciences
Qualitative Data Exchange Format (QuDEx)	Qualitative Social Sciences
Ecological Metadata Language (EML)	Ecology
Darwin Core	Biodiversity
MIxS	Genomics
IEDA Marine Geoscience Data System metadata form	Marine Geoscience
Directory Interchange Format (DIF)	Earth Sciences
Service Entry Resource Format (SERF)	Earth Sciences
IEDA System for Earth Sample Registration metadata templates	Earth Science
Digital Library for Earth System Education (DLESE)	Earth Science Education
Content Standard for Digital Geospatial Metadata (CSDGM)	Geographic
General Transit Feed Specification (GTFS)	Public Transportation
NEES metadata requirements	Earthquake Engineering
W3C Data on the Web Best Practices	General

FINDINGS

- Researchers invest more effort in documenting their data when they expect to share their data.
- Researchers do not use standards or community practices when creating documentation (and often, they aren't aware of any).
- Researchers feel their articles should be comprehensive enough to act as metadata.
- Researchers find it difficult to document unspoken assumptions and tacit knowledge.
- Metadata and documentation needs are incredibly diverse.

FINDINGS – COMMON METADATA NEEDS

- Title
- Description/Abstract
- Data Creator(s)
- Contributor(s)
- Organization
- Depositor
- Sponsor
- Keywords
- Copyright/License
- Embargo
- Language
- File Information
- Last Modified
- Related Publications
- Object of Study/Unit of Analysis
- Characteristics of Object of Study/Unit of Analysis
- Experimental Design/Setup
- Environmental or Experimental Conditions
- Time information/Time Period Covered
- Geographic Information/Place of Data Collection
- Date of Data Collection
- Methods
- Data Analysis
- Attribute or Code Definitions
- Project Description
- Project Name
- Research Design
- Sampling Methods/Protocol
- Code or scripts used in analysis
- Software
- Data Source/Source of data or samples
- Additional Information

FINDINGS – DISTINCT METADATA NEEDS

Economics

- Data Collection Instrument
- Known Limitations
- Descriptive Statistics
- Conceptual Framework
- Sample Size
- Variable Definitions

Interactive Computing

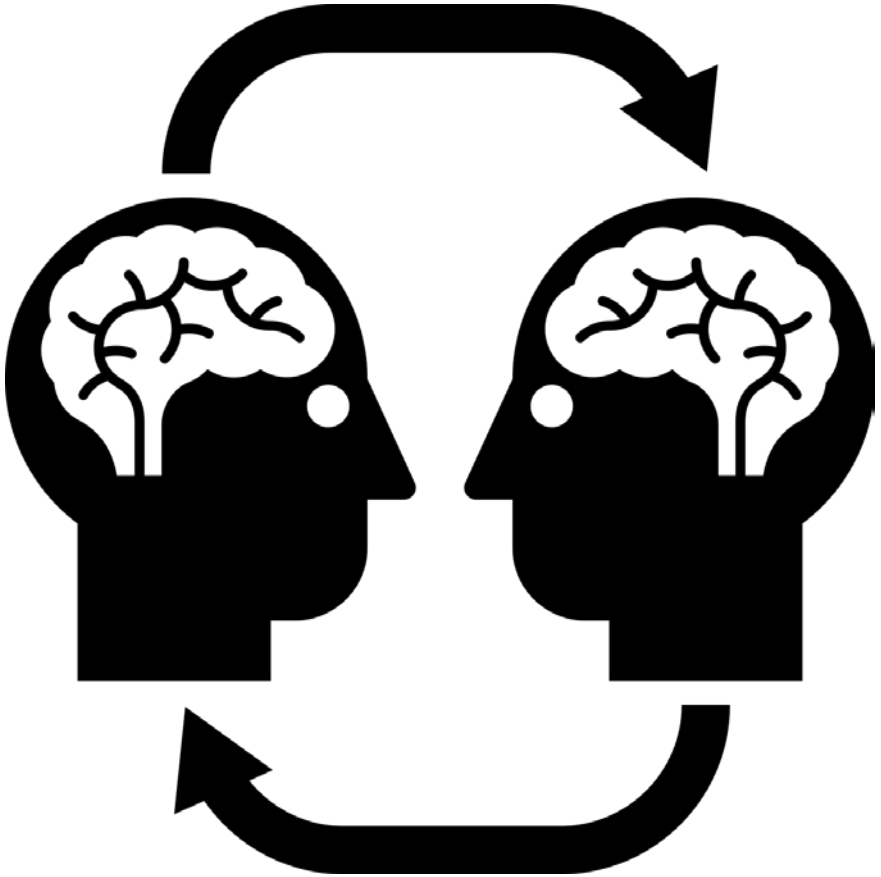
- Data Collection Instrument
- Accuracy and Quality Information
- Conceptual Framework
- Sample Size
- Variable Definitions

Civil & Environmental Engineering

- Accuracy and Quality Information
- Related manuals, user guides
- Equipment
- Taxonomic Information
- Standards used and level of compliance
- Definitions

FINDINGS

Begin to build community census about expectations for documentation and metadata



NEXT STEPS

- Explore needs of other Schools at Georgia Tech
- Explore differences in data types (qualitative vs. quantitative; simulation vs. experimental)
- Create web-form to collect information and create “README.txt”
- Evaluate what types of metadata we can support in structured metadata records
- Ask additional researchers to review current templates

REFERENCES

Brown KM, Burk LM, Henagan LM, Noor MAF (2004) Data from: A test of the chromosomal rearrangement model of speciation in Drosophila pseudoobscura. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.1150

Chao, T. (2015). Mapping methods metadata for research data. *International Journal of Digital Curation*, 10(1), 82-94. doi:10.2218/ijdc.v10i1.347

Dryad. (2015). Frequently Asked Questions. Dryad. Retrieved April 12, 2015 from http://datadryad.org/pages/faq.

University of Virginia (2015). Datasets. University of Virginia Library. Retrieved April 12, 2015 from https://pages.shanti.virginia.edu/libra/datasets/.