DATA WITH DONUTS:
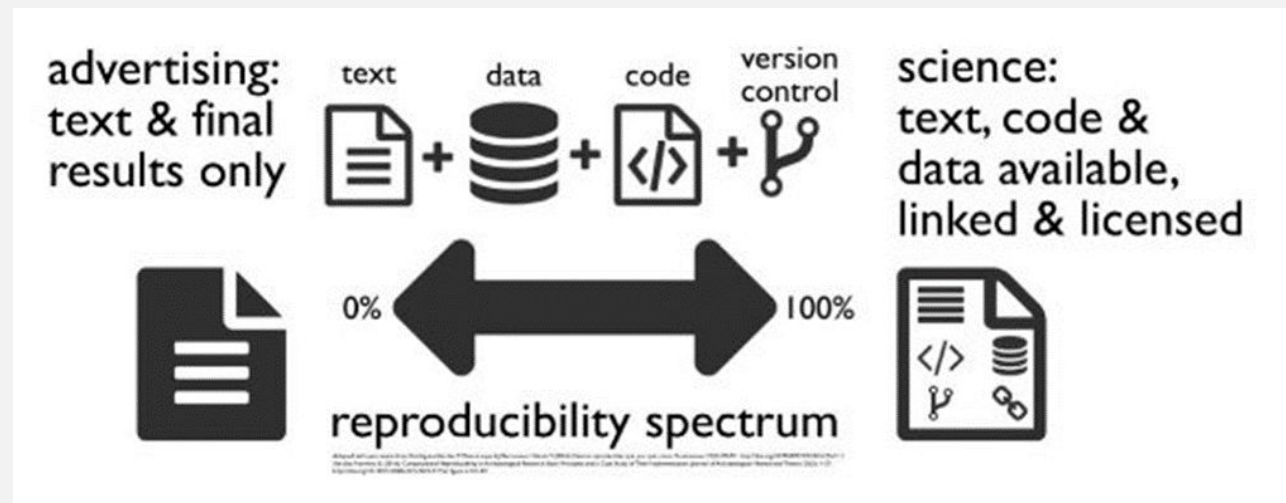# FINDING, ACCESSING AND CITING DATA

September 29, 2017

Kathy Hart Weimer

khw2@rice.edu

DATA USE AND PUBLISHING SPECTRUM

Publication of the data and / or code used for the analysis greatly improves reproducibility and supports open science.

DATA SOURCES - GOVERNMENT

- Data.gov
  - All US Federal Agencies, variety of themes, apps
  - Some state/city level data
- United Nations – data.un.org
  - 35 databases, including topics of: crime, energy, education, environment, food, gender, health, labor, population, refugees, tourism, trade
- And many others – http://libguides.rice.edu/data_resources

DATA SOURCES - REPOSITORIES BY DISCIPLINE

- •Re3data.org (hosted by DataCite)
  - > 1,500 worldwide research data repositories

- Nature – http://www.nature.com/sdata/policies/repositories
  - Nature Publishing Group → Scientific Data, an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets
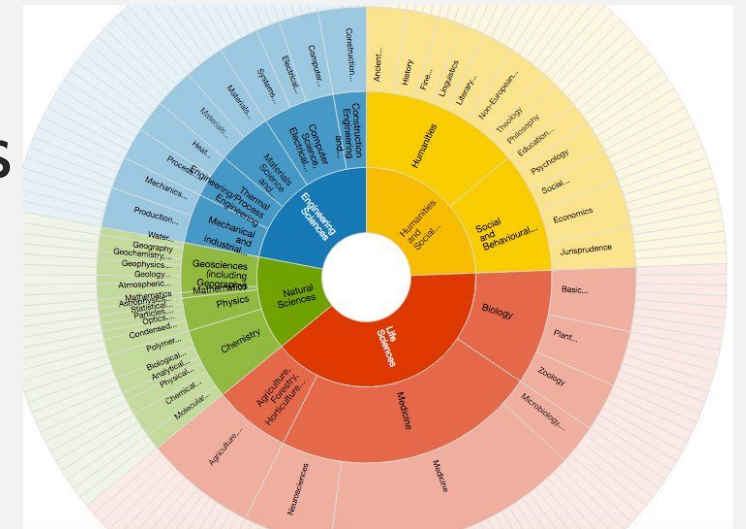
DATA SOURCES : SUBSCRIPTION OR MEMBERSHIP

- Dataverse / Figshare / Dryad
  - General, multidisciplinary, deposit and download
- ICPSR – Inter-University Consortium for Political and Social Research
- Social Explorer
  - Social, demographic, business data
- ProQuest Statistical Insight
  - US & International data from the 1970s to today

CITING DATA

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

Preamble from *Force11*
(Future of Research Communications and e-Scholarship)

FORCE11 PRINCIPLES

1. Importance – Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

2. Credit and Attribution – Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

3. Evidence – In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

4. Unique Identification – A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

5. Access – Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6. Persistence – Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

7. Specificity and Verifiability – Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

8. Interoperability and Flexibility – Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities

GENERAL GUIDES:

- MLA - no specific guidance, but core elements →

- APA - New in 6th ed. - data sets - see 7.08

- Data providers often provide a citation style - check their websites!

1 Author.
2 Title of source.
3 Title of container,
4 Other contributors,
5 Version,
6 Number,
7 Publisher,
8 Publication date,
9 Location.

RECOMMENDED CONTENT:

1. Author/Principal Investigator
2. Year of Publication
3. Title of the Data Source
4. Edition/Version Number
5. Format of the Data Source (e.g. [Computer File], [CD-ROM], [Online], etc.)
6. Producer of the Data Source
7. Distributor of the Data Source
8. Identifier or permanent URL for the Data Source

ORNL CITATION EXAMPLE

- Gu J.J., E.A. Smith, and H.J. Cooper. 2006. LBA-ECO CD-07 GOES-8 L4 Gridded Surface Ration and Rain Rate for Amazonia: 1999. Data Set. Available on-line [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/831

DRYAD EXAMPLE (ACCOMPANYING A PUBLICATION)

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters doi:10.5061/dryad.75nv22qj

ICPSR EXAMPLE

Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release [Computer file]. ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849

CONTACT US!

- Kathy Hart Weimer
- Head, Kelley Center
  - khw2@rice.edu

- Research Data Management Team
  - researchdata@rice.edu

Next Data & Donuts– Oct. 13

Storing, Backing Up and Archiving Data