Hands-on Activity 8: Data Citation

Authors: Matt Mayernik (National Center for Atmospheric Research Library), Gail Steinhart (Cornell University Library)

Background Lecture: Lesson 8: Data Citation

Objectives: Students will understand the rationale for data citation and the issues involved. They will understand what information is necessary to cite a data set, and consider when citation or authorship are appropriate when reusing data created by others.

Outcomes: Students will (1) consider the current state of data citation practice as it relates to the Joint Declaration of Data Citation Principles, (2) identify factors to consider when making decisions about citation or authorship when reusing a data set, (3) identify components of a data citation, and (4) create a data citation based on an existing metadata record.

Time Needed: In a classroom setting, 30-45 minutes for discussion of readings in class, 30 minutes for data citation exercise in class. 45 minutes outside of class (for key readings).

URLs: http://www.dataone.org/sites/all/documents/L08_DataCitation_20160922.pptx

Additional Files Needed: none

Key Readings:

Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. Martone, M. (ed.). San Diego, CA: FORCE11; 2014. https://www.force11.org/datacitation.

Duke, C. S. & Porter, J. H. The Ethics of Data Sharing and Reuse in Biology. BioScience 63, 483–489 (2013). http://dx.doi.org/10.1525/bio.2013.63.6.10

Notes for Instructors:

This exercise will engage students in a discussion of the importance and ethics of data citation (based on the key readings), and give them hands-on practice with reading and creating data citations. Instructors may choose to do one or both exercises.

Discussion of key readings

Have the students read the key readings prior to class. In class, have students discuss the questions listed below (see Student instructions).

Discussion question for Joint Declaration of Data Citation Principles:

The Data Citation Principles are meant to promote very general best practices related to data citation, both in terms of research practice and technical infrastructure. Which of the principles are largely accessible and implementable now, and which are more difficult or aspirational? Why? Additional questions for each principle follow below. *There aren't strictly right or wrong answers in this discussion, and the landscape will continue to evolve, but some possible responses are included below.*

1. Importance: What professional norms and practices, both of individuals and of institutions or organizations, support or undermine the idea that data are legitimate and citable products of research? How? *Possible answers include promotion and tenure criteria at universities (data may or may not be valued as scholarly contributions), research funders policies (they may encourage demonstrating the availability of data sets created with prior funding, data in CVs, and require sharing data), publishers may encourage or require data sharing and citation, data centers may provide recommended citations for data sets.*
2. Credit and attribution: Credit and attribution of more traditional types of research products is an established norm and practice; is extending this practice to include data a simple and natural thing to do? Why or why not? *On the surface, yes, but questions of what constitutes an authorship role may be an issue (see Duke and Porter paper), note also some of the possible responses to question 1.*
3. Evidence: Citing literature to support claims is also an established practice; is extending this practice to include data a simple and natural thing to do? Why or why not? *Not all data are in a data center or repository and formally citable (see Duke and Porter paper); authors may not be familiar with data citation practices.*
4. Unique identification: Is it always possible for a data creator to obtain a persistent identifier for their data set? Why or why not? *Repository or data center may or may not supply a persistent identifier, data creator may or may not have access to an identifier service to create one themselves; identifiers must be maintained in order to continue to work.*
5. Access: In practice, do data citations always provide direct access to the data set? Why or why not? *Sometimes - data centers only provide metadata and contact information and data must be obtained from the data creator or another party.*
6. Persistence: In practice, do data citations (and metadata) persist beyond the lifespan of the data set? Should they? Why or why not? *Persistence is the*

responsibility of the organization holding the data; organizations may or may not be long-lived, and plans for transfer or responsibility for a data set may or may not be in place. Individual researchers or research groups may have valid reasons for not making data available indefinitely. Data sets may have multiple versions; data creators may have valid reasons for wanting to make only a particular version (the most recent, for example) available.

7. Specificity and verifiability: Why is it important to be able to create and maintain specific and verifiable references to data sets, portions of data sets, or versions of data sets? What are some potential challenges to doing so? *Specific and verifiable citations support reproducible and repeatable research and analyses, especially when an analysis is conducted with a subset of a larger data set, but the technical infrastructure often does not readily support this level of citation.*
8. Interoperability and flexibility: What are some of the different stakeholder groups whose practices may influence the ability to support interoperability across citation standards and styles? *Journal publishers, data centers and repositories, funders, professional societies, and authors.*

Discussion question for Duke and Porter paper:

What are three factors when considering whether acknowledgement, formal citation, or co-authorship is the most appropriate way to provide attribution to the creator of a data set used in a publication?

1. Status and condition of data set: Data shared informally between researchers may not be formally citable - it may lack a persistent identifier, adequate metadata, and may not otherwise be publicly available. Informal acknowledgement is appropriate in this case.
2. Journal guidelines for authorship are appropriate to consider (but frequently do not address the contributions of data creators).
3. Usage rights, licensing statements, and other formal or informal conditions of use for the data set.
4. Importance of a data set to the overall analysis and publication.
5. Novelty of the data set.
6. Availability and willingness of data creator to fulfill other responsibilities of authorship.
7.

Also in class, have the students practice reading and writing data citations by completing the Citation exercises worksheet.

Student Instructions:

Read the two key readings before class.

Discussion question for Joint Declaration of Data Citation Principles:

The Data Citation Principles are meant to promote very general best practices related to data citation, both in terms of research practice and technical infrastructure. Which of the principles are largely accessible and implementable now, and which are more difficult or aspirational? Why?

1. Importance: What professional norms and practices, both of individuals and of institutions or organizations, support or undermine the idea that data are legitimate and citable products of research? How?

2. Credit and attribution: Credit and attribution of more traditional types of research products is an established norm and practice; is extending this practice to include data a simple and natural thing to do? Why or why not?

3. Evidence: Citing literature to support claims is also an established practice; is extending this practice to include data a simple and natural thing to do? Why or why not?

4. Unique identification: Is it always possible for a data creator obtain a persistent identifier for their data set? Why or why not?

5. Access: In practice, do data citations always provide direct access to the data set? Why or why not?

6. Persistence: In practice, do data citations (and metadata) persist beyond the lifespan of the data set? Should they? Why or why not?

7. Specificity and verifiability: Why is it important to be able to create and maintain specific and verifiable references to data sets, portions of data sets, or versions of data sets? What are some potential challenges to doing so?

8. Interoperability and flexibility: What are some of the different stakeholder groups whose practices may influence the ability to support interoperability across citation standards and styles?

Discussion question for Duke and Porter paper:

What are three factors when considering whether acknowledgement, formal citation, or co-authorship is the most appropriate way to provide attribution to the creator of a data set used in a publication?

Citation exercises

These activities use real data citations and data set metadata records to help you understand the challenges and processes for creating data citations.

1. Identify the elements of a data citation - The following data citations are real examples taken from DataONE member node data centers. What information is provided in these citations? Pull the citations apart and describe each component. Do these citations appear to follow any data citation formats discussed in the module? If so, which one(s)?
    a. Ellison, Aaron; Bennett, Katherine (2009): Sarracenia Purpurea Prey Capture at Harvard Forest 2008. Long Term Ecological Research Network. http://dx.doi.org/10.6073/pasta/9a6105374adb15486b75cf621a2702dd
    b. Nepstad, D.C., E.A. Davidson, D. Markewitz, E.J.M. Carvalho, J.Q. Chambers, D. Ray, J.B. Guerrero, P. Lefebvre, L. Sternberg, M. Moreira, L. Barros, F.Y. Ishida, I. Tohlver, E.L. Belk, K. Kalif, and K. Schwalbe. 2012. LBA-ECO ND-30 Water Chemistry, Rainfall Exclusion, km 67, Tapajos National Forest. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. http://dx.doi.org/10.3334/ORNLDAAC/1131

2. Creating data citations - Create citations for the following data sets using the information found in the metadata records provided below. Choose a standard citation format described in the module.

    DATA SET #1

    author: Tognetti, Pedro Maximiliano
    author: Chaneton, Enrique Jose
    coverage.spatial: Inland Pampa
    coverage.spatial: Buenos Aires
    coverage.spatial: Argentina
    coverage.spatial: South America
    date.accessioned: 2014-09-16T18:41:59Z
    date.available: 2014-09-16T18:41:59Z
    date.issued: 2014-09-15
    identifier: doi:10.5061/dryad.46181
    description: 1. Native vegetation fragments embedded in anthropogenic landscapes are increasingly threatened by land-use intensification. Managing disturbance regimes and nutrient inputs may help maintain species diversity in such remnants. Yet it is unclear the extent to which changes in resource availability due to reduced capture by resident plants and/or increased supply rates may trigger native community disassembly and exotic invasions. 2. We examined how mowing disturbance and N fertilizer addition affected plant community recovery after a burning event in a remnant corridor of tussock pampa grassland in Argentina. The percentage cover and richness of native and exotic plant functional groups were monitored over four years. According to the 'fluctuating resource theory', we expected invasion to be highest when both light and N availability were increased simultaneously. 3. Mowing

delayed recovery by dominant C4 tussock grasses and promoted subordinate, native C3 grasses and exotic legumes, thus enhancing both native and exotic species richness. Fertilization induced a transient increase in native forbs but decreased total plant richness. Moreover, N addition to mowed grassland led to rapid invasion by short-lived exotic forbs, which were then replaced by exotic perennial grasses. Exotic grasses eventually spread across the grassland corridor, although at different rates depending on the treatment, and in parallel to a generalized decline in native species cover. 4. Synthesis and applications. Community disassembly patterns reflected differential responses of native and exotic functional groups to altered resource supply rates. Synergisms between canopy disturbances and N enrichment posed the greatest threat to preserving a pampa grassland remnant prone to invasion. Establishing buffer zones may be required to enhance the viability of corridor-like grassland remnants in agricultural landscapes.
subject: burning
subject: functional groups
subject: invasion
title: Data from: Community disassembly and invasion of remnant native grasslands under fluctuating resource supply
data center: Dryad Digital Repository

DATA SET #2

Title: Soil properties and nutrient concentrations by depth from the Anaktuvuk River Fire site in 2011
Author: M. S. Bret-Harte, Michelle C. Mack, G. Shaver, J. Laundre
Point of Contact: Michelle Mack
Description: Below ground soil bulk density, carbon and nitrogen was measured at various depth increments in mineral and organic soil layers at three sites at and around the Anaktuvuk River Burn: severely burned, moderately burned and unburned. This data corresponds with the aboveground biomass and root biomass data files: 2011ARF_AbvgroundBiomassCN, 2011ARF_RootBiomassCN_byDepth, 2011ARF_RootBiomassCN_byQuad, 2011ARF_RootBiomassCN_byQuad.
Time Coverage: Jul 24, 2011 - Jul 28, 2011
Northernmost Latitude: 68.99
Southernmost Latitude: 68.99
Westernmost Longitude: -150.28
Easternmost Longitude: -150.28
Science Keywords:
Land Surface > Soils > Soil Bulk Density
Land Surface > Soils > Nitrogen

Land Surface > Soils > Carbon
Location(s): United States Of America > Alaska
Platform(s): Field Survey
Instrument(s): Chemical Meters/Analyzers > CHN ANALYZERS > Carbon, Hydrogen, Nitrogen Analyzers (Carbon, Hydrogen, Nitrogen Analyzers)
Data Format(s): Microsoft Excel spreadsheet
Language: English
Date Created: 2013-12-05 16:31:39
Date Last Updated: 2015-02-05 10:49:56
Data Center: Advanced Cooperative Arctic Data and Information Service
URL: https://www.aoncadis.org/dataset/2011ARF_SoilCN_byDepth.2.html