Curating context and use: Pulling scientific workflows into the repository

Andrea Thomer - Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science - University of Illinois at Urbana-Champaign

Abstract: Though data curation is often discussed as a discrete process (e.g. Higgins 2008), in reality it is highly dependent on scientific workflows. Yet, a data curator's understanding of a dataset is typically limited to her experience of it within a repository: separate from its context of production, and separate from its use. Here, we explore ways data curators can collect metadata about context and use via two projects at the National Center for Atmospheric Researcher (NCAR).

Context: An audit of CESM model output metadata

How much curation of climate model output can be done without direct collaboration with climate modelers?

Prediction: Though there is a community-developed controlled vocabulary for climate model output (the NetCDF Climate and Forecast Metadata Convention (CF)), few modelers will use it. We expect that some variable names will be easily cross-walked to CF, but others will require input from modelers for curation.

Design: collected header data from 5 different files and tallied (per file):
-number of "standard_terms" in CF (below)
-number of terms crosswalkable to CF
-number of terms not crosswalkable to CF

```
double ilev(ilev) ;
     ilev:long_name = "hybrid level at interfaces (1000*(A+B))" ;
     ilev:units = "level" ;
     ilev:positive = "down" ;
     ilev:standard_name = "atmosphere_hybrid_sigma_pressure_coordinate" ;
     ilev:formula_terms = "a: hyai b: hybi p0: P0 ps: PS" ;
```

Preliminary results:
< 4% of terms variables contained "standard_names" in CF
~ 17% could have been crosswalked

Follow up work: extended audit of NetCDF headers

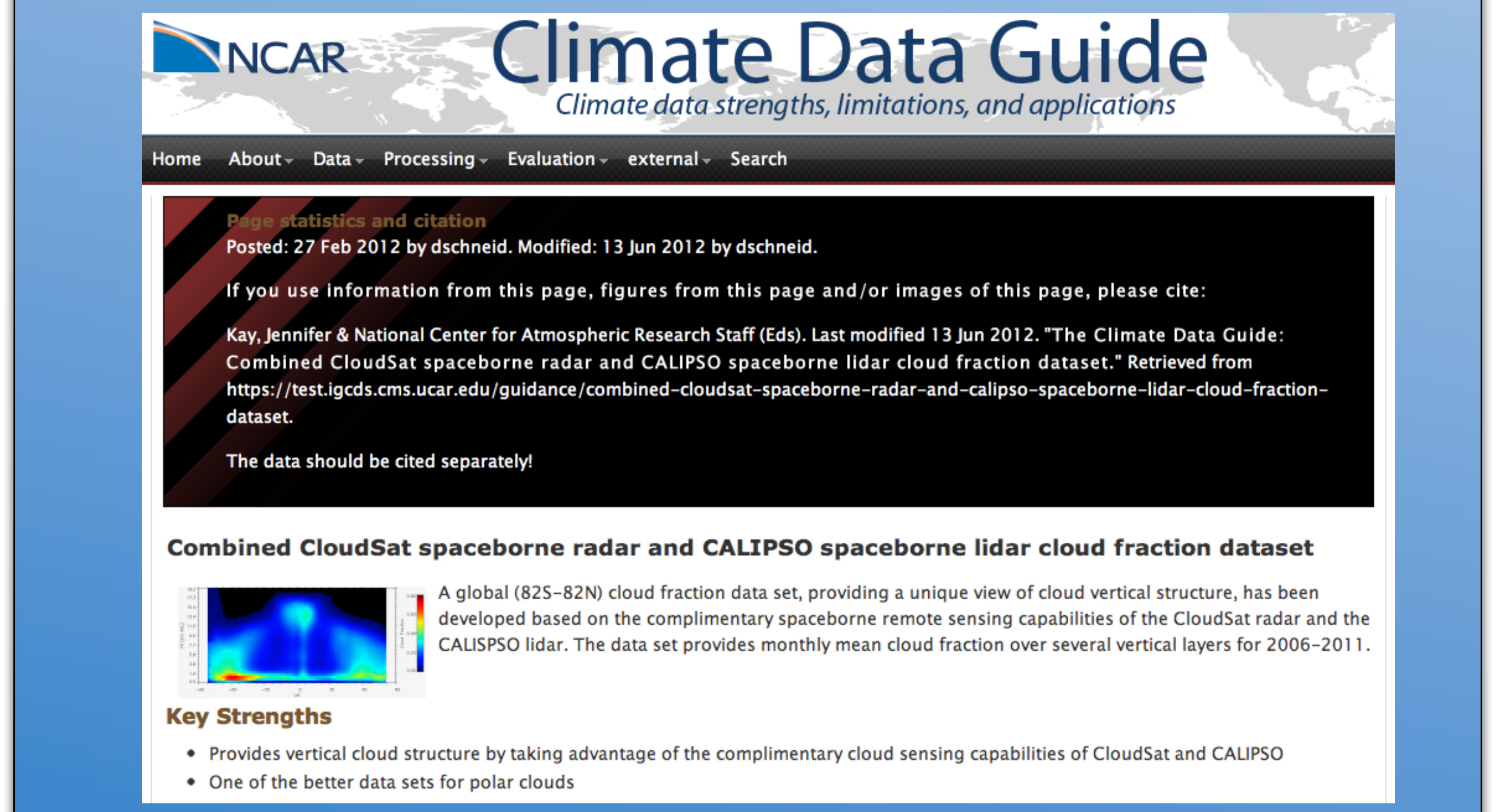
Use: Encouraging CDG contribution by encouraging citation

How do we encourage contributions to the Climate Data Guide (CDG)? – (http://climatedataguide.ucar.edu/)

Prediction: Encouraging CDG users to cite the guide will, in turn, encourage citation-hungry scientists to contribute to the Guide.

Design: via Drupal, added PHP-enabled "blocks" at the top of each page containing an automatically generated, easily copy-and-pasteable "suggested citation" in which contributors are named as editors (below).

Follow up work: tracking citations of the CDG, looking for indications that there is a correlation between citability and contribution.



Page statistics and citation
Posted: 27 Feb 2012 by dschneid. Modified: 13 Jun 2012 by dschneid.

If you use information from this page, figures from this page and/or images of this page, please cite:

Kay, Jennifer & National Center for Atmospheric Research Staff (Eds). Last modified 13 Jun 2012. "The Climate Data Guide: Combined CloudSat spaceborne radar and CALIPSO spaceborne lidar cloud fraction dataset." Retrieved from https://test.igcds.cms.ucar.edu/guidance/combined-cloudsat-spaceborne-radar-and-calipso-spaceborne-lidar-cloud-fraction-dataset.

The data should be cited separately!

Combined CloudSat spaceborne radar and CALIPSO spaceborne lidar cloud fraction dataset

A global (82S-82N) cloud fraction data set, providing a unique view of cloud vertical structure, has been developed based on the complimentary spaceborne remote sensing capabilities of the CloudSat radar and the CALISPSO lidar. The data set provides monthly mean cloud fraction over several vertical layers for 2006-2011.

Key Strengths

- Provides vertical cloud structure by taking advantage of the complimentary cloud sensing capabilities of CloudSat and CALIPSO
- One of the better data sets for polar clouds

In both projects:

- We can take advantage of shared community and organizational structure to pull data about production and use into an otherwise isolated cyberinfrastructure.
- Curatorial input from scientists must be inferred if not directly obtained.
- Future work: can extrinsic motivators designed to increase curatorial input from scientists (citations, enforced use of standards) turn into intrinsic habits?

This work was supervised by Gary Strand, David Schneider, and Dennis Shea (NCAR), and funded by the Institute of Museum and Library Services as part of the Data Curation Education in Research Centers program.

References:
Higgins, S. (2008). The DCC Curation Lifecycle Model. The International Journal of Digital Curation, 3(1), 134-140.