# Recognizing Diversity of Contributions: Framing Data Attribution and Acknowledgement

Chung-Yi (Sophie) Hou – <a href="mailto:hou@Illinois.edu">hou@Illinois.edu</a>
Advisor: Matt Mayernik (National Center for Atmospheric Research)

May, 2015





# **Introduction / Background**

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Specialization in Data Curation



\*Data Curation Education in Research Centers Program



# **Research Interest**

```
collection * components * curation * cyberinfrastructure * cycle * data*
development * different * digital * dissertation* ecosystem*elements * factors * focus
* information * infrastructure * interoperability * layer * life *
                   lifecycle * management * metaphor *
   model*play *preservation* protest * questions * report * research * role *
       science * social * Stages * system * technical * think *understanding * work
   collection * components * curation * cyberinfrastructure * cycle * data*
development * different * digital * dissertation * eCOSYSTEM*elements * factors * focus
  * information * infrastructure * interoperability * layer * life *
                   lifecycle * management * metaphor *
   model play * preservation * geolect * questions * report * research * role *
      science * social * Stages* system * technical * think *understanding * work
```

Kouper, I. (2012, August 20). Data Curation Week: Just a Beginning. [Web Blog Image] Retrieved from <a href="http://connect.clir.org/CLIR/Blogs/BlogViewer/?BlogKey=13436e05-25bc-4fe1-b506-592c02e6b9f7">http://connect.clir.org/CLIR/Blogs/BlogViewer/?BlogKey=13436e05-25bc-4fe1-b506-592c02e6b9f7</a>



# **Observation**

What about citation, acknowledgement, attribution?



# **Question - 1**

 What encouraged and supported the current citation style?

**Example of a Current Citation Style Format\*:** 

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

\*DataCite: https://www.datacite.org/services/cite-your-data.html



# **Question - 2**

 Is the current citation style suitable and sufficient as the datasets become more collaborative and extensive in the diversity of skill sets and expertise?

### **Example of A Current Citation:**

Rife, D. L., Pinto, J. O., Monaghan, A. J., Davis, C. A., and Hannan, J. R. (2014): NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Dataset.

http://dx.doi.org/10.5065/D6M32STK. Accessed 20 Apr 2015.



# **Question - 3**

 What will the next generation attribution system look like?

- Continuation of current method?
- Complete new model?
- Hybrid styles?



# **Inspiration**

Data Science Journal, Volume 12, 10 February 2013

### IS DATA PUBLICATION THE RIGHT METAPHOR?

M A Parsons1 \* and P A Fox2

\*1 National Snow and Ice Data Center, University of Colorado, UCB449, Boulder, CO 80309

Email: parsons.mark@gmail.com

<sup>2</sup>Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180

Email: pfox@cs.rpi.edu

### **ABSTRACT**

International attention to scientific data continues to grow. Opportunities emerge to re-visit long-standing approaches to managing data and to critically examine new capabilities. We describe the cognitive importance of metaphor. We describe several metaphors for managing, sharing, and stewarding data and examine their strengths and weaknesses. We particularly question the applicability of a "publication" approach to making data broadly available. Our preliminary conclusions are that no one metaphor satisfies enough key data system attributes and that multiple metaphors need to co-exist in support of a healthy data ecosystem. We close with proposed research questions and a call for continued discussion.

**Keywords**: Data publication, Data system design, Data citation, Semantic Web, Data quality, Data preservation, Cyberinfrastructure



# **Case Study**

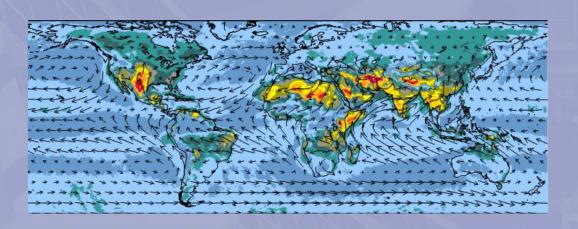


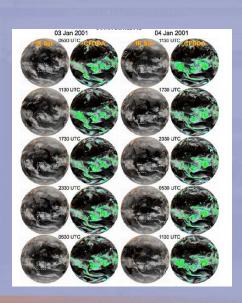




# **Case Study - Continued**

- Climate Four-Dimensional Data Assimilation (CFDDA) dataset is:
  - Created over a two-year period
  - High temporal (hourly from 1985 to 2005)
  - High spatial (40km horizontal grid with 0.4 degree grid and 28 vertical level)







# **Preparation**

Verify Data Quality

Consistency Validation Compliance Harvest Metadata Descriptions

> Tool Format Content

Document Provenance Information

The Data
Curation
Profiles



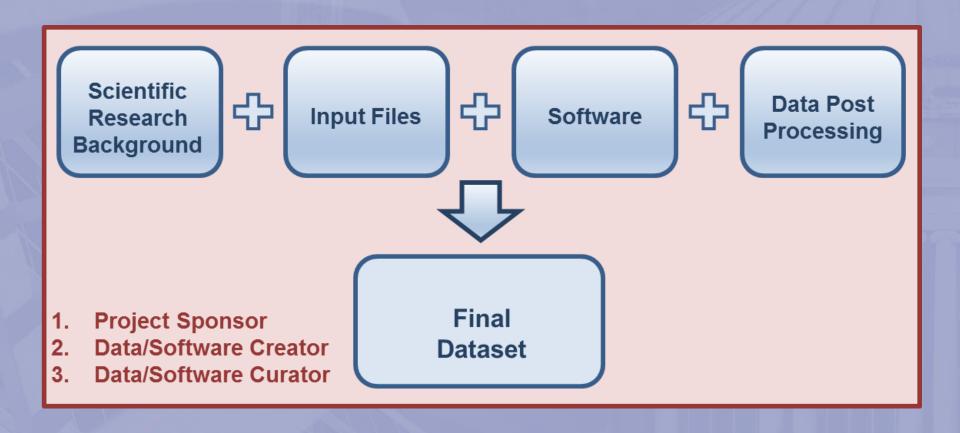
# **Preparation - Continued**

Rife, D. L., Pinto, J. O., Monaghan, A. J., Davis, C. A., and Hannan, J. R. (2014): NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Dataset. http://dx.doi.org/10.5065/D6M32STK. Accessed 20 Apr 2015.

- Scientists who were cited were not satisfied with the limitation of the citation.
  - They would like additional teammates and those who supported them to also be attributed and acknowledged.
- Curation process also revealed additional roles and responsibilities that enable the production and management of the dataset.



# **Methodology / Process**





# <u>Methodology / Process - Continued</u>

Item Title	Organization	Individuals	Rationale
MDVBlend and MDVCombine	Project Sponsor: - N. A.	Project Sponsor: - N. A.	- These tools are part of the MDV data analysis tools suite. These tools are used for
	Software Creator: - Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	Software Creator: - N. A.	"stitching the hemispheres together" or to produce the composite meshes for final CFDDA dataset.
	Software Curator: - Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	Software Curator: - N. A.	
MDVtonetcdf	Project Sponsor: - N. A.	Project Sponsor: - N. A.	- This tools is part of the MDV data analysis tools suite. It is used to convert CFDDA data
	Software Creator: - Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	Software Creator: - N. A.	format from MDV to netCDF.
	Software Curator: - Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	Software Curator: - N. A.	
Climate Data Operation (CDO)	Project Sponsor: - N. A.	Project Sponsor: - N. A.	- CDO is open source and released under the terms of the GNU General Public License
	Software Creator: - N. A.	Software Creator: - N. A.	v2. It is essential for performing statistical analysis of netCDF file. The attribution information
	Software Hosting Site: - Max-Planck-Institut fur Meteorologie	Software Curator: - Cedrick Ansorge - Kameswar Rao Modali - Ralf Quast - Luis Kornblueh - Ralf Mueller	is based on CDO's home page: https://code.zmaw.de/projects/c do
		- Uwe Schulzweida	

# **Result**

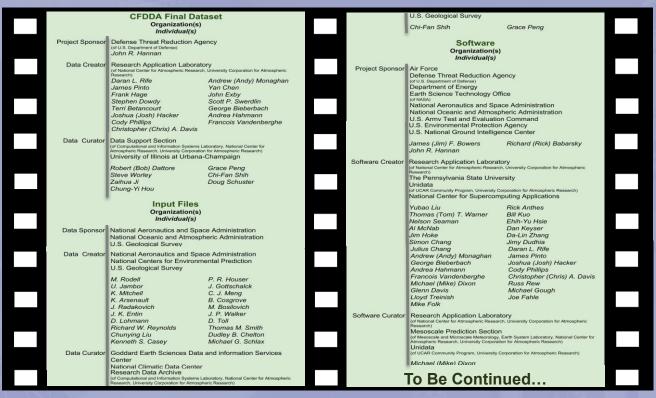
# A total of 26 unique organizations and 103 unique individuals were identified.

VS.

1 organization and 5 authors in the current citation.



### **Result - Continued**



Vs.

Rife, D. L., Pinto, J. O., Monaghan, A. J., Davis, C. A., & Hannan, J. R. (2014). NCAR
Global Climate Four-

Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis. [Dataset]. Retrieved from http://rda.ucar.edu/datasets/ds604.0



### **The Alternatives**



IMDb.com, Inc. (2014). *The Imitation Game*. Retrieved from http://www.imdb.com/title/tt2084970/?ref\_=nv\_sr\_1

Develop "Project Workbook" to document 9 key aspects of the dataset?

- Project overview
- Initiation plan and SSR
- 3. Project scope and risks
- Management procedures
- 5. Data descriptions
- 6. Process descriptions
- Team correspondence
- 8. Project Charter
- 9. Project schedule

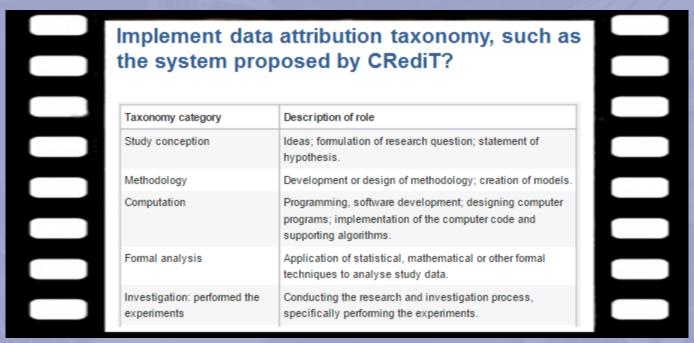
Online copies of data dictionary, diagrams, schedules, reports, etc.



Hoffer, J. A., George, J. F., & Valacich, J. S. (2014). Managing the Information Systems Project. In Modern Systems Analysis and Design Seventh Edition (pp. 81). Essex, England: Pearson Education Limited



### **The Alternatives - Continued**



CRediT. (2014). *Proposed Terms*. Retrieved from <a href="http://credit.casrai.org/">http://credit.casrai.org/</a>
Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014, April 16). *Publishing: Credit where credit is due*. Retrieved from <a href="http://www.nature.com/news/publishing-credit-where-credit-is-due-1.15033">http://www.nature.com/news/publishing-credit-where-credit-is-due-1.15033</a>



# **Lessons Learned**

 Use major milestones/key phases of the project to organize the contribution areas.

- Leverage existing taxonomies or controlled vocabularies to enhance content consistency.
- Time is of the essence when it comes to creating and maintaining attribution/acknowledgement.



### Crediting a Climate Model Dataset Like a Movie? - A Case Study in Data Attribution

Chung-Yi Hou (hou@Illinois.edu)<sup>1</sup>, Terri Betancourt (terrib@ucar.edu)<sup>2</sup>, Matthew Mayernik (mayernik@ucar.edu)<sup>3</sup>

1 - Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign; 2 - Research Applications Laboratory, National Center for Atmospheric Research (NCAR); 3 - NCAR Library, National Center for Atmospheric Research (NCAR)

### Introduction

As climate models' data volumes, format types, and sources increase rapidly with the invention and improvement of science, climate model datasets are becoming more complex to manage as well.

One of the significant management challenges is pulling apart the individual contributions of specific people and organizations within large complex projects. This is important both for 1) assigning responsibility and accountability for scientific work, and 2) giving professional credit to individuals (e.g. hiring, promotion, and tenure) who work within such large projects.

Analogous to acknowledging the different roles and responsibilities in movie credits, the methodology developed in this study that was used to identify and map out the relationships among the organizations and individuals who had contributed to the dataset could provide a useful framework for constructing dataset attribution in general.

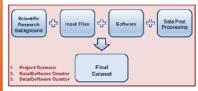
### Research Objectives

Using the NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40km Reanalysis1 as the dataset for the case study, the case study aimed to:

- Identify the unique individuals and organizations who had contributed directly to the production of the CFDDA dataset.
- · Model the individuals and organization attribution in the style of

### Method

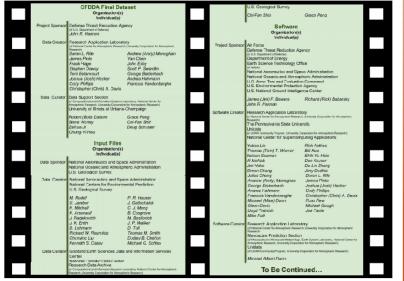
Preparation: Based on the metadata documentation and provenance information compiled by the authors during the curation phase of the CFDDA dataset, the authors identified the following 5 categories and 3 roles that participated in the production of the CFDDA dataset.



Data Collection: The authors sequentially and systematically analyzed the metadata documentation and the provenance information for each of the 5 categories in order to identified the unique individuals and organizations who fit the roles and had contributed directly to the production of the CFDDA dataset.

A total of 26 unique organizations and 103 unique individuals were identified

The following shows a sample of the attribution in the movie credits style



(Attributions for the "Data Post Processing" and the "Scientific Research Background" categories are not included in the above sample.)

### Lessons Learned

Using a non-publication metaphor, the study showed that attributing datasets like movie credits would help clarify the roles involved in producing and

Particularly, significantly more organizations and individuals could be acknowledged when the roles and types of contribution are expanded beyond primary authors as one might be accustomed to see in a traditional journal publication setting, such as the citation shown in the Reference section of this

However, due to the complexity and the length of history for datatsets, especially those that have been worked on over a long time duration, the following lessons learned from the case study could be used to help with organizing and constructing the attribution:

- Integrate and manage the process of documenting the roles and responsibilities of contributing organizations and individuals as part of the dataset project life cycle
- This includes determining the key information, such as contribution type, name, contact information, job title, etc., that should be collected For software or tools that have several revisions, set the "depth" or the number of revisions that have clear and direct contribution to the production of the dataset.
- Keep the definitions and formats of the data attribution consistent.
- · If missing information are detected, collect and complete the information for all the categories of the data attribution.

### **Alternative Attribution Options**



Option 1: Build online database that collects

### Option 2: Implement data attribution xonomy, such as the system proposed by

Taxonory rategrey	Description of trial	
Study surrogation	Non, hardeles d'estes hyenite, plateres d' hypothèsic	
Metrodriogr	Development a design of methodology, creation of model	
Computation	Programming, software development, cestigning computer programs, implementation of the computer code and supposing algorithms.	
roma asaysis	Application or issession, mathematica or other formal federal (see a new postary dista.)	
investigation; percented the	Conducting the research and investigation provess,	

### Option 3: Develop "Project Workbook" to document 9 key aspects of the dataset

- Project overview
- Initiation plan and SSR . Project scope and risks
- Management procedures
- Data descriptions
- Team correspondence
- Project Charter
- . Project schedule

### **Future Work**

- Software is one area that is currently not as strong in terms of curation. This affects the amount of available information to maintain data attribution over time
- The impact of cloud computing on the practices of data curation, and therefore data attribution, requires further study.

The authors would like to thank the Data Curation Education in Research Centers [DCERC] project, funded by the Institute of Museumand Library Services [RE-02-19-0304-19], for inspiring and providing the learning opportunity.

CRediT. (2014). Proposed Terms. Retrieved from <a href="http://credit.com/microscopy">http://credit.com/microscopy</a> Alson I., Scott, J., Brand, A., Have, M., & Alman, M. (2014). Spril 16). Publishing Credit where credit is due. Retri

### ∞ LIBRARY AND INFORMATION SCIENCE







http://www.dcc.ac.uk/sites/default/files/documents/IDCC15/175 Creatingaclimatemodel.pdf



# **Future Work / Next Steps**

- How to define "contribution"?
- Is the use of the phases of the project lifecycle a viable method for organizing contribution areas?
- What are some of the other data types that should be evaluated?
- In terms of contributing roles, what are other resources in addition to CRediT to consider?
- What is the impact of change in granularity of the attribution/acknowledgement content?
- How to implement the framework?



### **Thank You!**

Questions and Comments?

Please feel welcome to contact me at hou@Illinois.edu

