Recognizing Diversity of Contributions: Framing Data Attribution and Acknowledgement

Chung-Yi (Sophie) Hou – hou@Illinois.edu
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

DCERC

+ NCAR

INSTITUTE of MUSEUM and LIBRARY SERVICES

DataONE

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
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:
What will the next generation attribution system look like?

- Continuation of current method?
- Complete new model?
- Hybrid styles?
IS DATA PUBLICATION THE RIGHT METAPHOR?

M A Parsons\textsuperscript{1,*} and P A Fox\textsuperscript{2}

\textsuperscript{1}National Snow and Ice Data Center, University of Colorado, UCB449, Boulder, CO 80309
Email: parsons.mark@gmail.com
\textsuperscript{2}Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8\textsuperscript{th} 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 - 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

- 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**

1. **Project Sponsor**
2. **Data/Software Creator**
3. **Data/Software Curator**

*Diagram showing the process of combining Scientific Research Background, Input Files, Software, and Data Post Processing to create a Final Dataset.*
<table>
<thead>
<tr>
<th>Item Title</th>
<th>Project Sponsor</th>
<th>Organization</th>
<th>Individuals</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDVBlend and MDVCombine</td>
<td>Project Sponsor:</td>
<td>- N.A.</td>
<td>Project Sponsor:</td>
<td>- These tools are part of the MDV data analysis tools suite. These tools are used for “stitching the hemispheres together” or to produce the composite meshes for final CFDDA dataset.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software Creator: - Research Application Laboratory (RAL), National Center</td>
<td>Software Creator: - N.A.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software Curator: - Research Application Laboratory (RAL), National Center</td>
<td>Software Curator: - N.A.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDVtonetcdf</td>
<td>Project Sponsor:</td>
<td>- N.A.</td>
<td>Project Sponsor:</td>
<td>- This tool is part of the MDV data analysis tools suite. It is used to convert CFDDA data format from MDV to netCDF.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software Creator: - Research Application Laboratory (RAL), National Center</td>
<td>Software Creator: - N.A.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software Curator: - Research Application Laboratory (RAL), National Center</td>
<td>Software Curator: - N.A.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate Data Operation</td>
<td>Project Sponsor:</td>
<td>- N.A.</td>
<td>Project Sponsor:</td>
<td>- CDO is open source and released under the terms of the GNU General Public License v2. It is essential for performing statistical analysis of netCDF file. The attribution information is based on CDO’s home page:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Software Hosting Site: - Max-Planck-Institut fur Meteorologie</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Result

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

vs.

1 organization and 5 authors in the current citation.
The Alternatives

Build online database that collects attribution and dataset information like IMDb.com?


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

1. Project overview
2. Initiation plan and SSR
3. Project scope and risks
4. Management procedures
5. Data descriptions
6. Process descriptions
7. Team correspondence
8. Project Charter
9. Project schedule
Implement data attribution taxonomy, such as the system proposed by CRediT?

<table>
<thead>
<tr>
<th>Taxonomy category</th>
<th>Description of role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study conception</td>
<td>Ideas; formulation of research question; statement of hypothesis.</td>
</tr>
<tr>
<td>Methodology</td>
<td>Development or design of methodology; creation of models.</td>
</tr>
<tr>
<td>Computation</td>
<td>Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms.</td>
</tr>
<tr>
<td>Formal analysis</td>
<td>Application of statistical, mathematical or other formal techniques to analyse study data.</td>
</tr>
<tr>
<td>Investigation: performed the experiments</td>
<td>Conducting the research and investigation process, specifically performing the experiments.</td>
</tr>
</tbody>
</table>


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), Terri Betancourt (terrib@ucar.edu), Matthew Mayernik (mayernik@ucar.edu)
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 modelers’ data volumes, formats, and sources increase rapidly with the invent 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 having the accountability and transparency for scientific work, and 2) giving professional credit to individuals (e.g. training, promotion, and tenure) who work on such large projects.

An analogous task is identifying the different roles and responsibilities in movie credits. The methodology developed in this study was used to identify and map out the relationships among the organizations and individuals who contributed to the dataset, to provide a useful framework for constructing dataset attribution in general.

Research Objectives

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

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

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 5 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 identify the unique individuals and organizations who fit the roles and had contributed directly to the production of the CFDDA dataset.

Results

A total of 20 unique organizations and 133 unique individuals were identified.

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

Future Work

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

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

ILLCNOIS.EDU
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