IJDC | Peer-Reviewed Paper

Recognizing the Diversity of Contributions: A Case Study for Framing Attribution and Acknowledgement for Scientific Data

Chung-Yi Hou
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Matthew Mayernik
UCAR/NCAR Library
National Center for Atmospheric Research
University Corporation for Atmospheric Research

Abstract

As scientific data volumes, format types, and sources increase rapidly with the invention and improvement of scientific capabilities, the resulting 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 for two aspects: 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. This paper aims to review the extant practice of data attribution and how it may be improved. Through a case study of creating a detailed attribution record for a climate model dataset, the paper evaluates the strengths and weaknesses of the current data attribution method and proposes an alternative attribution framework accordingly. The paper concludes by demonstrating that, analogous to acknowledging the different roles and responsibilities shown in movie credits, the methodology developed in the study could be used in general to identify and map out the relationships among the organizations and individuals who had contributed to a dataset. As a result, the framework could be applied to create data attribution for other dataset types beyond climate model datasets.

Received 21 April 2015 ~ *Accepted* 26 June 2016

Correspondence should be addressed to Chung-Yi Hou, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000. Email: hou@ucar.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v11i1.357

Introduction

The advent of the eScience era – a concept "denoting the use of digital technology for solving scientific problems" (Greenberg, White, Carrier and Scherle, 2009) – has brought the proliferation of data in a variety of types and formats. While there were concerns regarding "data deluge," which could lead to information pathologies such as information overload, information anxiety, and information avoidance (Bawden and Robinson, 2009), others saw the potential for the advancement of science through multi-disciplinary data integration and synthesis.

One of the science disciplines that has experienced the advantages of data proliferation is the climate sciences. Combining data from different measurement types to produce climate model datasets with high temporal and spatial resolution is essential to the conduct of climate science (Edwards, 2010). Climate model datasets combined with heterogeneous observational data provide opportunities to examine previously studied climate phenomena with new perspectives, and to discover new climate patterns. However, in order to produce climate model datasets that could offer high resolution and detailed information, each climate model dataset project needs diverse amount of resources, including financial, infrastructural and human. Additionally, to enable the projects' success, each of these resources requires careful coordination. The management and acknowledgement of the expertise, knowledge and skills that contributed to such projects can have significant impact in boosting team morale, establishing accountability and earning promotions. The need to provide attribution to the project contributors is further emphasized when creating author lists on papers that publish project results.

Traditionally, journal publication citations have been the baseline method for providing public acknowledgement. Additionally, acknowledgement for data could be provided within journals either as in-text references to data's authors and the associated analyses or as a special "thank you" under the specified acknowledgement section. However, these citation and acknowledgement formats were developed for journals published mainly in paper form. In other words, the traditional citations offered only the basic information, primarily focused on the first author or principal investigator. The traditional citations were also aimed to provide identifications predominantly for the articles published in journals, and did not provide detail coverage of the comprehensive contributions involved with the creation and maintenance of the data, which often provided the basis and support for the published journal articles.

With new information types, especially in digital formats such as climate model datasets, the contribution types involved in producing and managing the datasets can be extensive. The traditional journal citation and acknowledgement formats might no longer be sufficient to provide the additional desired acknowledgement of different roles and responsibilities for the various contributions. As a result, in order to provide the full context of the effort involved in producing and managing climate model datasets, it is important to explore a new framework for providing attribution to scientific work so that in depth and committed participations from diverse sources of skills and expertise can be encouraged and recognized.

Background and Rationale

Recent research based on the research life cycle model has shown how data management and data sharing should be integrated more closely within research processes (Tenopir et al., 2011). The concept of a data life cycle model has informed the understanding of key concepts related to documenting metadata to provide data descriptions, using common data formats to facilitate data interoperability, and creating repositories to allow data sharing. Subsequently, these concepts have also gained attention and traction for further developments. However, the practices of data citation and acknowledgement have remained consistent with past practices, in which detailed attribution for data-related tasks are rare (Parsons, Duerr and Minster, 2010). Additionally, although the term 'data citation' might sometimes be used interchangeably with data attribution, there are subtle yet crucial differences: citation more often refers to the mechanism by which one makes references to other entities while attribution is more closely associated with the notion of recognizing contribution than the identification of authorship (Borgman, 2012). Since datasets and other information types undergo different creation and development processes than journal articles, applying traditional publication citations to datasets may result in inadequate acknowledgement to the diverse skills and expertise involved.

In the case of the NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40km Reanalysis dataset, citation became an important focal point for discussion during the curation phase of the dataset. In particular, as a part of the curation process, the curation team, led by the authors, wanted to provide a citation for the dataset so that it would be easy for the end users to reference the dataset. Sharing the philosophies advocated by DataCite[1], the curation team also believed that by providing and directly associating a citation with the CFDDA dataset, the readily available citation information would encourage the end users to cite the dataset. Consequently, the citation would also help in promoting the practices of recognizing and rewarding data producer/provider. In addition, the use of citation would allow the dataset's impact to be traced, and therefore promote reuse and verification of the dataset. As a result, the team discussed the method and format in which the citation should be assigned to the datasets. During this process, the curation team interviewed the CFDDA dataset's project manager and the science team, and reviewed the project's original records. In the end, the citation that was developed for the CFDDA[2] dataset followed primarily the format recommended by DataCite, but the citation content could also be restructured to accommodate the formats recommended by American Geophysical Union, American Meteorological Society, Federation of Earth Science Information Partners, and Geoscience Data Journal. However, it was during these discussions and based on the documents examined that the curation team, including the lead scientists of the CFDDA dataset, realized that the traditional journal publication citations could not adequately represent all the individuals and organizations who had contributed to the dataset. Alternatively stated, since the roles

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

2 CFDDA dataset's final 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.

 doi:10.2218/ijdc.v11i1.357

and responsibilities that contributed to the creation and maintenance of the dataset extended well beyond the principal investigators, the traditional journal citations did not reflect the full scope of the effort that was required to produce and manage the dataset. Consequently, the authors were inspired to review alternate frameworks that could present appropriately the contributing members of the dataset.

Literature Review

Numerous studies have analyzed the concept of 'authorship' within scientific and academic institutions. These studies have shown how notions of authorship are highly contextual, and contain many social and political nuances (Biagioli and Galison, 2003). Many of these studies focus on the ways that authorship norms and practices have evolved in the 20th and 21st centuries. Author lists on scholarly research articles have been increasing in length for a number of decades as collaborative science has become common (Greene, 2007). One outcome of large collaborative scientific projects is that individuals face significant challenges in gaining recognition for their contributions among dozens or hundreds of project members (e.g. Birnholtz, 2006; Galison, 2003). In large collaborations, such 'hyperauthorship' (Cronin, 2001) can make the precise contributions of each individual hard to distinguish from the outside. Author lists themselves are highly variable across domains and journals, with relatively few research articles listing authors using simple alphabetic lists (Waltman, 2012). In addition, early career or contract researchers may have little communication with their supervisors about authorship criteria, or little power to influence the decisions related to such criteria (Tarnow, 1999; Tilbury, 2007).

These trends are similar across the academic domains, but have manifested differently in individual research areas (Cronin, Shaw and La Barre, 2003; 2004). High-energy physics is known to have some of the largest author lists on individual papers, with most author lists numbering above ten and a large number of papers listing hundreds of authors (Tarnow, 2002). The biomedical fields have also grappled extensively with the definition of authorship, with numerous articles being published on the topic in the past twenty years (Steen, 2013). The preoccupation with authorship in the biomedical fields extends beyond just the length of the author lists. Many discussions also focus on the authorial status (or lack thereof) of students, technicians, and post-docs in laboratory research, and on honorary authorship practices, where a lab leader will be listed on each article produced by lab members regardless of the leader's actual contribution to any individual paper (Flanagin, et al., 1998; CSE Task Force on Authorship, 2000). A result of these debates is that definitions of authorship are now commonly found in the author guidelines of biomedical journal publishers (Osborne and Holland, 2009).

Various schemes have been proposed with the goal of reducing the ambiguity of authorship on large scale collaborations, most involving the attribution of authorship based on the idea of 'contributors,' or attribution with specific reference to the role each individual played in the research process (ex. Rennie, Yank and Emanuel, 1997; Paneth, 1998; Davenport and Cronin, 2001). These schemes often draw an analogy to the credit lists shown at the beginning or ending of movies. The scrolling of the end credits make the contributions of each individual visible, even if the vast majority of names and their associated roles mean little to the viewing audience (Conley, 2003).

Scientific research projects likewise often involve individuals with a far wider range of roles than that of 'author'.

Given the centrality of authorship to the career tracks of scholarly researchers, a range of proposals have been made to formalize the contributor roles involved in scholarly work. In 2001, Davenport and Cronin (2001) proposed that academic articles use an Extensible Markup Language (XML) mark-up to indicate the individuals who contributed to the production of papers. Their proposal included 19 contributor types, ranging from conception/design of a study, to collecting data, and reviewing/approving the final manuscript. While no implementations directly took Davenport and Cronin's proposal forward, a number of more recent initiatives cover similar ground. The DataCite organization, which provides organizations with the capability to assign Digital Object Identifiers (DOIs) to research resources, created a typology of 'contributors' that can be designated within metadata associated with DOIs. Within version 3.1 of the DataCite schema, the typology lists 21 specific contributor types, including types for individual people and organizations. The typology is implemented in an XML schema as a 'contributorType' attribute for a 'contributor' element (DataCite, 2014). Another notable initiative is developing a Contributor Role Taxonomy under the name CRediT[3]. CRediT, developed via the Consortia Advancing Standards in Research Administration Information (CASRAI) and the US-based National Information Standards Organization (NISO), includes 14 contributor roles. The motivation for the development of the CRediT taxonomy was to encourage standardization of contributor role designations within publication workflows, such as within journal submission or manuscript management systems (Brand, Allen, Altman, Hlava, and Scott, 2015).

All of these contributor typologies focus on providing more granularities to research attribution practices, similar to the idea of spelling out individual contributions to the production of movies that are listed in movie credits. However, more detailed attribution designations do not necessarily reduce the problem of attribution. The compilation of movie credits, for example, involves detailed social and political negotiation and mediation. The designation of screenwriting credits for movies often involves arbitration by the Writers Guild of America, which has the final say over screenwriting credits for most movies (Welkos, 1998). Attribution systems for scholarly research must likewise navigate and accommodate a variety of social and political institutions.

Method

The production of the CFDDA dataset was originally completed in 2009. After the dataset's completion, the final version of the dataset consisted of 183,960 files and was nearly 27TB in volume. The CFDDA dataset was made available to the project sponsor, and had been used by researchers to model weather and climate patterns. Due to the initial agreement with the project sponsor, the CFDDA dataset had also been managed by the project sponsor's partner, Research Applications Laboratory (RAL) at the National Center for Atmospheric Research (NCAR), on servers that were only accessible by project team members. During the summer of 2014, after the project sponsor had released the CFDDA for public use, the dataset was curated and deposited by the authors in collaboration with CFDDA experts at RAL into NCAR

3 CRediT: http://credit.casrai.org/

 doi:10.2218/ijdc.v11i1.357

Computational and Information Systems Laboratory's (CISL) Research Data Archive (RDA). The CFDDA dataset currently is considered to be one of the highest temporal and spatial resolution reanalysis dataset that is available in the world. Therefore, the purpose of the curation work was to facilitate the dataset's public accessibility, understandability and usability via CISL RDA, so that communities beyond NCAR RAL could also benefit from the scientific merits of the dataset.

During the curation process the dataset's data format and content were verified, metadata and documentation were gathered and compiled, and a recording of the dataset's provenance was produced. As a result of this curation effort, the authors realized the large, diverse set of contributions that enabled the creation of the CFDDA dataset were not well reflected by the formal CFDDA citation noted above. Specifically, as part of the data curation process, the authors used the guidance provided by Data Curation Profiles[4] to interview the CFDDA dataset's project manager and lead scientists. The Data Curation Profile included 13 modules that were designed to assist with the examination and reporting of a project's characteristics in a systematic manner. Among the different team members, the project manager and the lead scientists were selected for the interview process due to their comprehensive involvement in the planning, execution, and management of the CFDDA project. Consequently, they had in depth project knowledge not only in the technical areas but also in terms of personnel. During these interviews and through reviewing the modules in the Data Curation Profile, the project manager and the lead scientists were able to share extensive details regarding the project life cycle and the related activities. Combined with careful reviews of the project's original documents, the authors were able to identify and summarize five key phases of the project life cycle: Scientific Research Background, Input Files, Software, Data Post Processing, and Final Dataset. Additionally, the authors noticed that there were a wide variety of job titles that were involved with the project, and three roles had existed consistently in all five key phases: Project Sponsor, Data/Software Creator, and Data/Software Curator. Table 1 and Table 2 show respectively the definition assigned by the authors to each of these five areas and the three roles. After presenting the interview summary and findings to the project manager and the lead scientists, the team realized that specific individuals and organizations involved in each of these five phases could be further analyzed, identified and grouped based on the three repeating roles in order to demonstrate a more comprehensive documentation of the project history. By compiling the details of these contributions, the diversity of individuals and organizations that was required and was indeed involved in producing, managing and curating such complex dataset could also be made available for others to review in order to improve their understanding of the dataset.

After the authors had defined the five areas and the three roles for contribution documentation, the authors proceeded to analyze sequentially and systematically the metadata documentation and the provenance information in order to determine the specific information relating to the contributing individuals and organizations. During this analysis process the authors also included the CFDDA project manager and the lead scientists in the review and verification process. In particular, the project manager and the lead scientists helped ensure that the contributing roles and responsibilities of the identified individuals and organizations, as well as the project phase that they were categorized under, were all correct to the actual occurrence of the project. Subsequently, the contribution content had been fully reviewed and approved by the core CFDDA project team members. Likewise, their feedback also

4 Data Curation Profiles: http://datacurationprofiles.org/

helped in structuring and confirming the final contribution framework. Finally, once identified and confirmed, the unique individuals and organizations that fitted the definitions of contributors to the production and maintenance of the datasets were documented and tabulated.

After all the documents and information had been reviewed, the finalized relationships and roles of the individuals and organizations based on the definitions were mapped using a movie-credit like format.

In the case of the CFDDA project, the project was well documented, and the project manager as well as the lead scientists were all motivated in helping to record the contributions of their team members. Specifically, the project manager and the lead scientists expressed appreciation for the team and were interested in recognizing and officially documenting the full effort that was involved with the project. As a result, the project manager and the lead scientists were consistently proactive in providing project details. This allowed the authors to encounter minimum challenges when inquiring and obtaining pertinent and accurate contribution information. It would be important to note that if the project had not preserved its documents completely or if the key members of the team were not available to provide such project information, it would be much more difficult to collect and verify the contribution information retrospectively.

Table 1. Definitions of Contribution Areas.

Contribution Area	Definition
Scientific Research Background	The individual or organization who was responsible for establishing the scientific foundation and knowledge base for the production of the dataset.
Input Files	The individual or organization who was responsible for providing and setting up the initial input conditions for the production of the dataset.
Software	The individual or organization who was responsible for providing, setting up, and maintaining the analysis environment for the production of the dataset.
Data Post Processing	The individual or organization who was responsible for analyzing, reviewing, and synthesizing the components needed for the production of the dataset.
Final Dataset	The individual or organization who was responsible for performing verification and quality control, as well as producing the final deliverable version of the dataset.

 doi:10.2218/ijdc.v11i1.357

Table 2. Definitions of Contribution Roles.

Contribution Role	Definition
Project Sponsor	The individual or organization who was responsible for providing the resources, including financial, human, and infrastructural, to enable the production of the dataset.
Data/Software Creator	The individual or organization who was responsible for the primary construction of the data and software components required for the production of the dataset.
Data/Software Curator	The individual or organization who was responsible for providing long term management and stewardship of the data and software components required for the production of the dataset.

Results

This section outlines how a total of 26 unique organizations and 103 unique individuals were identified as having significant contributing roles in the creation and management of the CFDDA dataset. These organizations and individuals comprised those who had direct or indirect contributions to the CFDDA dataset. Direct contributions included those that actually worked on the CFDDA dataset or had immediate impact on the dataset, and indirect contributions included those that helped support and influence the availability of the resources that underlay the production of the CDFFA dataset.

To provide an example of the results, Table 3 shows the detailed analysis performed to determine the contributors to the Data Post Processing portion of the CFDDA project. It shows the key organizations and individuals who contributed to the five different key data post processing software and tools that were used for the CFDDA dataset. The right hand column shows the rationale for why these organizations and individuals were determined to be relevant contributors.

Table 3. Contributions to CFDDA's Data Post Processing.

Item Title	Organization	Individuals	Rationale
MDVBlend and MDVCombine	**Project Sponsor[5]:** N/A **Software Creator:** Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University	**Project Sponsor:** N/A **Software Creator:** N/A	These tools are part of the MDV data analysis tools suite. These tools are used for "stitching the hemispheres together" or to

5 The National Center for Atmospheric Research and is sponsored by the National Science Foundation. RAL and CISL are laboratories within NCAR that benefits from NSF funding through the use of physical facilities, computational resources, and other services from which these projects receive overall support.

Item Title	Organization	Individuals	Rationale
	Corporation for Atmospheric Research (UCAR) **Software Curator:** Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	**Software Curator:** N/A	produce the composite meshes for final CFDDA dataset.
MDVtonetcdf	**Project Sponsor[5]:** N/A **Software Creator:** Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR) **Software Curator:** - Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	**Project Sponsor:** N/A **Software Creator:** N/A **Software Curator:** N/A	This tools is part of the MDV data analysis tools suite. It is used to convert CFDDA data format from MDV to netCDF.
Climate Data Operation (CDO)	**Project Sponsor:** N/A **Software Creator:** N/A **Software Hosting Site:** Max-Planck-Institut fur Meteorologie	**Project Sponsor:** N/A **Software Creator:** N/A **Software Curator:** Cedrick Ansorge Kameswar Rao Modali Ralf Quast Luis Kornblueh Ralf Mueller Uwe Schulzweida	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

 doi:10.2218/ijdc.v11i1.357

Item Title	Organization	Individuals	Rationale
			home page[6]
netCDF Operator (NCO)	**Project Sponsor:** N/A **Software Creator:** Department of Earth System Science, University of California, Irvine **Software Hosting Site:** SourceForge.net	**Project Sponsor:** N/A **Software Creator:** - Charles S. Zender **Software Curator:** Charles (Charlie) Zender Henry Butowsky Wenshan Wang	NCO is used to perform tasks that CDO is unable to do. The attribution information is based on NCO's home page[7] and citation guidance[8].
NCAR Command Language (NCL)	**Project Sponsor[5]:** N/A **Software Creator:** Visualization & Enabling Technologies Section (VETS), Computational and Information Systems Laboratory (CISL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR) **Software Curator:** Visualization & Enabling Technologies Section (VETS), Computational and Information Systems Laboratory (CISL), National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)	**Project Sponsor:** N/A **Software Creator:** N/A **Software Curator:** Mary Haley Dennis Shea Adam Phillips Cindy Bruyere Sherrie Fredrick	NCL is used to visualize the CFDDA data. The attribution information is based on NCL's home page[9], the citation guidance[10], and the contributor section of NCL's DOI Metadata[11].

The same analysis technique was used on the other four contribution areas (note: the full analysis tables are omitted for length). Tables 4 and 5 show the final count for each of the direct and indirect contributing areas and roles defined. Please note that

6 CDO: https://code.zmaw.de/projects/cdo
7 NCO: http://nco.sourceforge.net/
8 NCO Citation Guidelines: http://nco.sourceforge.net/nco.html#Citation
9 NCL: http://www.ncl.ucar.edu/
10 NCL Citation Guidelines: http://www.ncl.ucar.edu/FAQ/#misc_001
11 NCL DOI Metadata: http://data.datacite.org/10.5065/D6WD3XH5

the totals shown in these tables provide the cumulative counts for areas identified; as a result, the total values have not discounted organizations or individuals who participated in more than one contribution area.

Table 4. Total Count for Each Direct Contributing Areas and Roles Defined.

Direct Contribution	Organization	Individuals
Input Files: Data Creator	3	20
Software (RT-FDDA): Software Creator	1	1
Software (RT-FDDA): Software Curator	1	N/A[12]
Software (MM5): Software Creator	2	12
Software (Obs-Nudging): Software Curator	1	N/A[12]
Software (CFDDA): Project Sponsor	2	2
Software (CFDDA): Software Creator	1	9
Software (CFDDA): Software Curator	1	N/A[12]
Software (MDV): Software Creator	1	1
Software (MDV): Software Curator	1	1
Data Post Processing (MDVBlend, MDVCombine, and MDVtonetcdf): Software Creator	1	N/A[12]
Final Dataset: Project Sponsor	1	1
Final Dataset: Data Creator	1	15
Final Dataset: Data Curator	2	7
Total	**18**	**69**

12 No verifiable or confirmed contribution information could be found.

 doi:10.2218/ijdc.v11i1.357

Table 5. Total Count for Each Indirect Contributing Areas and Roles Defined.

Indirect Contribution	Organization	Individuals
Scientific Background	N/A[12]	53
Input Files: Project Sponsor	3	N/A[12]
Input Files: Data Curator	4	2
Software (RT-FDDA): Project Sponsor	1	1
Software (MM5): Project Sponsor	4	N/A[12]
Software (MM5): Software Curator	1	N/A[12]
Software (Obs-Nudging): Project Sponsor	N/A[12]	N/A[12]
Software (Obs-Nudging): Software Creator	N/A[12]	N/A[12]
Software (MDV): Project Sponsor	N/A[12]	N/A[12]
Software (netCDF): Project Sponsor	2	N/A[12]
oftware (netCDF): Software Creator	2	6
Software (netCDF): Software Curator	1	N/A[12]
Data Post Processing (MDVBlend, MDVCombine, and MDVtonetcdf): Project Sponsor	N/A[12]	N/A[12]
Data Post Processing (MDVBlend, MDVCombine, and MDVtonetcdf): Software Curator	1	N/A[12]
Data Post Processing (Climate Data Operation, netCDF Operator, and NCAR Command Language): Project Sponsor	N/A[12]	N/A[12]
Data Post Processing (Climate Data Operation, netCDF Operator, and NCAR Command Language): Software Creator	2	1
Data Post Processing (Climate Data Operation, netCDF Operator, and NCAR Command Language): Software Curator	3	14
Total	**24**	**77**

It is important to note that the list of organizations and individuals identified did not consist comprehensively of all the organizations and individuals that the authors were able to find to have traceable connections to the dataset. For instance, there were 71 additional scholarly journals that were also cited in an internal report to show their impact and contribution to the creation and development of the CFDDA dataset. However, these journals could not be fully analyzed for this study due to the confidentiality of the internal report. Similarly for the scientific journals that were publicly cited by the science team, since only the traditional journal citations were available, only the author names indicated by the citations could be identified for certain. Hence, additional associations that might also have contributed, such as the authors' affiliated organizations and institutions, were not identified.

Nevertheless, this study did determine accountable attributions based on the publicly known roles and contributions that had verifiable and confirmed connection to the production of the CFDDA dataset. As a result, the study demonstrated that though it remained challenging to define the complete scope and depth of attribution for a dataset, especially if significant portion of the attribution was documented only in traditional journal citations, it was still possible to improve the detail of attribution. Figure 1 shows the sample view of CFDDA dataset's attribution and acknowledgement content presented in style of a movie-credit format.

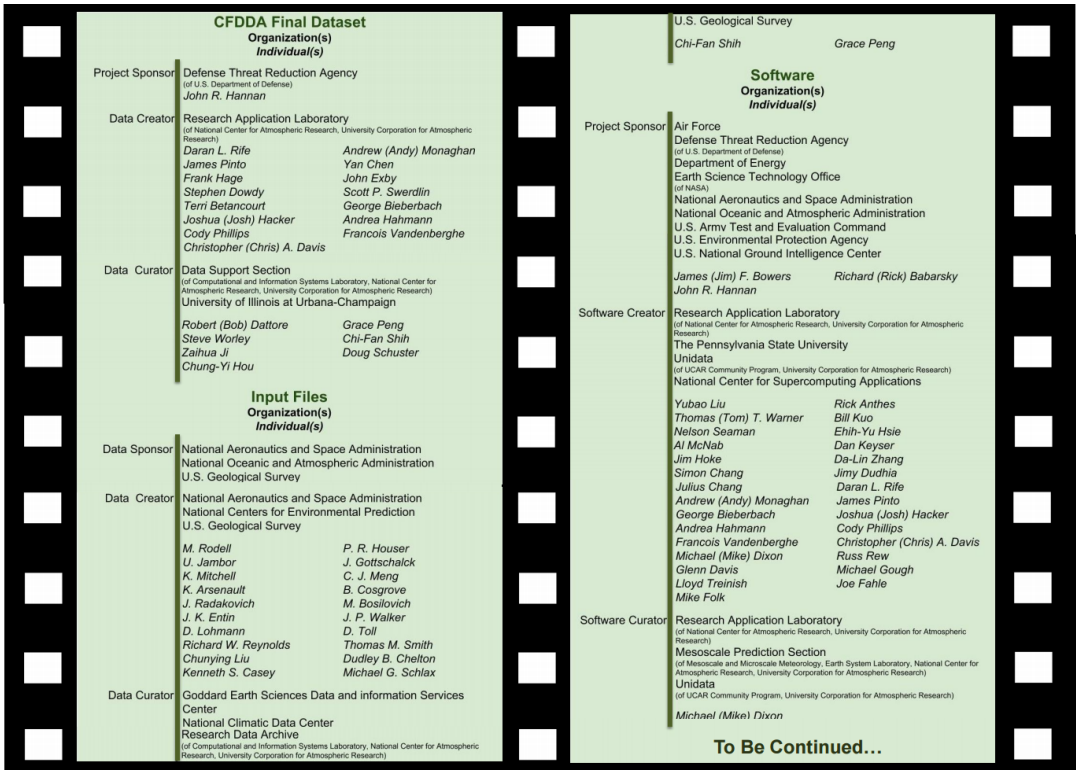

Figure 1. Sample of CFDDA Dataset's Attribution and Acknowledgement Content in Movie-Credit Format.

 doi:10.2218/ijdc.v11i1.357

Discussion

Overall, the case study aimed to highlight the importance of expanding attribution and acknowledgement to roles and responsibilities beyond primary authors or principal investigators. This would be applicable especially to scientific research projects that involved diverse skill sets and expertise, such as a climate model dataset. As the result of undergoing the process of compiling and formatting the attribution and acknowledgement content for the CFDDA dataset, additional observations were made that could serve as guidelines for generating attribution and acknowledge content for other scientific projects. These observations are summarized and discussed in the following sections.

Overall Framework Structure vs. Specific Detail Format

The case study was focused on providing a method for generating a framework to structure the attribution and acknowledgement content for a large, complex project. Specifically, while the major milestones or the key phases of the project life cycle could be helpful in identifying the main contribution areas and categories for attributing and acknowledging specific roles, such as in the case of the CFDDA project, other projects might have different details that could be used to define the contribution information, including the exact titles of roles given within the project. Consequently, this case study did not seek to standardize the format or syntax for recording the details of the attribution and acknowledgement content. Instead, the case study focused on the process of developing the overarching framework. This way, the general framework could be reused while the content format and syntax could be defined by any project to ensure that it would be appropriate to its local context.

Enhance Content Consistency through Standardization

Although the case study's aim was to demonstrate a framework for representing a project's diverse responsibilities and the relationships among the roles, this framework's consistency could be enhanced by integrating other standards or established practices. This could be especially helpful when providing detailed and extensive records of roles and responsibilities that would be shared with a wide community. For example, an online database, similar to the Internet Movie Database[13], could be used to publish and invite additional feedback for the attribution information. Developing a 'Project Workbook' based on the practice that would be familiar to systems managers (Hoffer, George and Valacic, 2014) could also be another option to record and maintain the attribution and acknowledgement content. Further, data attribution taxonomies, such as the system proposed by CRediT, could provide a standard set of definitions for describing roles and responsibilities that contributed to the production of a dataset.

When used in conjunction with the additional standards or best practices, the framework developed in the case study could potentially provide a more consistent structure, based on a project and data life cycle approach, for mapping out and relating the contribution areas and roles of a project. The resulting framework would encourage consistent use of well defined contribution areas and roles so that the terminologies and their uses could be clearly understood across different projects. As

13 Internet Movie Database: http://www.IMDb.com

a result, the framework should also support the use of best practices and a system of controlled vocabularies or a set of taxonomy, such as CRediT, so that the attribution and acknowledgment content could be adopted and understood by a wide community.

Leveraging Familiar Metaphors for Ease of Adoption

As the result of reviewing and documenting the role and responsibilities involved in the life cycle of the CFDDA dataset, the authors wanted to demonstrate a framework that could encompass multiple functions as well as outline the relationship among the different roles and responsibilities. Additionally, the authors wanted to select a framework that would be easy for the attribution and acknowledgement content to be visualized. Consequently, the motivation to choose a movie-credit style was to provide a familiar metaphor.

By referring to an easily relatable framework, the authors aimed to leverage the familiarity with movie credits in order to build quick understanding of the framework through mental association. Once the mental association was formed, it should also be easier to envision the possibility of expanding the roles and responsibilities for attribution beyond simple author lists. Such metaphors would help people to understand and construct concepts and communications (Parsons and Fox, 2013). Consequently, if data repositories or published papers present attribution information using a movie credit style, they could help to clarify how many different individuals and organizations played important roles in producing and managing datasets. Moreover, a movie credit-like presentation format could allow the realization that significantly more organizations and individuals could be acknowledged when the roles and types of contribution were expanded beyond primary authors. This contrast could especially be made when the movie credit-like presentation format was compared to the citations that one might be accustomed to see in a traditional journal publication setting, such as the CFDDA citation shown in Footnote 2 and the other citations in the References section of this paper.

Considerations for Implementation: Time is of the Essence

In order to adopt an attribution framework that could identify and maintain the diverse relationships and roles of contributions, it is important to evaluate and plan for the creation and the maintenance of the attribution and acknowledgement content before the start of a dataset project. Due to the complexity and the length of history for datasets, especially those that had been worked on over a long time duration, intricate relationships between the dataset project and its resources are common. Consequently, if the relevant information for the attribution and acknowledgement is not recorded in a timely manner, significant connections to many traceable contributions could be lost, or at best painstakingly identified at a later date. Likewise, in order to allow the attribution documentation to remain manageable over time, it is also crucial to strike a balance between defining the reasonable scope of the attribution and keeping as comprehensive as possible the acknowledgment of the roles and responsibilities types that had made contributions to a dataset project. As a result, the following lessons learned from the case study could be used to help when considering the organization and construction of an attribution/acknowledgement:

- Define and be consistent with the ‘terms and conditions’ of data attribution method/framework used.

 doi:10.2218/ijdc.v11i1.357

    - This includes determining the key information, such as contribution type, name, contact information, job title etc., that should be collected.

    - Once determined, the definitions and formats of the data attribution should also be kept consistent.

- Plan early for the resources needed to manage, organize, and store the information of data 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.

- In the case of the CFDDA project, the contribution information was reviewed and compiled based on the availability of the original project team members and documentations. If these resources did not exist or were not accessible, it would make the recording of the contribution information retrospectively very challenging, if not impossible. By including the documentation of the contributing roles and responsibilities as part of the project life cycle, it offers the advantage of each team member being able to help in providing his/her contributions as they occur. This could, therefore, also help share the task of contribution documentation and ensure the details as well as the appropriate context for the contribution are recorded properly. However, a post project review of the contribution information should also be conducted. This is so that any gaps or additional relevant clarifications could be added accordingly in order to complete or further enrich the documentation.

- 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.

- Allow scalability and extensibility in the framework, for example, if new contribution types are identified.

Areas for Further Exploration

In order to expand the study to understanding and development of data attribution further, a couple of open questions point toward possible next steps. First, how can data attribution be extended and applied to other data types, such as software? Many challenges related to attribution are similar across these different resource types, but the details may vary. Being able to test out the framework with different data types would allow the framework to be refined further, so that it could be adaptable as a general guideline. This framework was explored with two additional cases, software and dynamic datasets, in a separate paper (Hou and Mayernik, 2016). In addition, there could potentially be a synergy between data management and data attribution documentation. Data management planning guidance typically recommends reviewing the necessary tasks before starting a project and periodically re-evaluating the data management elements as the project is in process. However, given there are also various challenges associated with implementing data management plans for scientific projects, it would be important to understand how adding the documentation

of attribution information as part of the project's data management plan might further complicate the process. Furthermore, in terms of the implementation, what would be the most efficient method for assisting with the attribution documentation as well as sharing the attribution and acknowledgement content for reuse? In order to ensure that the attribution information could be captured and recorded in a timely manner, the documentation of contributions should not rely solely on manual effort and should be aided by tools that could help with the recording process. Likewise, once the attribution record is produced, any portion of the documentation should be quickly identifiable and referenceable, so that any different types of contributing areas and their associated roles and responsibilities can be easily reviewed and understood. Given the development and the growth of linked data and semantic web, constructing and applying an ontology in addition to creating an XML schema for the framework might be one possibility that could be evaluated. The authors chose to develop the framework through the creation of an initial XML schema. This schema has also been demonstrated with a set of associated case studies (Hou and Mayernik, 2015). Finally, it would be important to assess the impact of such contribution collection to the practices of data citation and attribution. Currently, the traditional citations place emphasis on a few people that are recognized to be the primary authors. With a more extensive recording of the contributing information, it should help provide an improved context and a holistic view of the contributing expertise and skill sets. Subsequently, there should also be opportunities for citations and attributions to be made for additional roles and responsibilities. As a result, it would valuable to determine if there might be a change or increase in crediting various categories of contributions and the purposes for these citations and attributions.

Conclusion

The volume of scientific data has been growing dramatically since the beginning of digital technology. The coming of the eScience age will continue the proliferation of data. While terminologies such as 'Big Data' and 'Data Deluge' could be construed with negative or intimidating connotations, the availability of diverse data could also have significant potential in assisting scientific advancement.

In the area of climate sciences, various data types are needed when creating high resolution climate model datasets. In addition to data, a variety of skilled team members are also required to facilitate the successful production of climate model datasets. The recognition of the diversity of research contributions will be significant for both professional and social reasons in the digital age. As a result, expanding attribution and acknowledgement frameworks beyond the traditional journal citations, which placed the emphasis on crediting the primary authors and principal investigators, could help in showcasing not only the range of resources, but also the different types of skill sets required to support and enable a scientific project.

Using the CFDDA dataset from NCAR as the basis for the attribution and acknowledgement framework case study, the authors were able to demonstrate that key phases or milestones in a project's life cycle could be used as the basic structure for organizing the contributing areas. In addition, by reviewing the project details with project team members, specific roles could be identified to help categorize the types of contributions made by the participating individuals and organizations. Furthermore, a movie credit model could be used to present the attribution and

acknowledgement structure and content to help others who might be outside of the immediate project to understand the relationships between the different contributing areas and roles. In the case of the CFDDA dataset, implementing the attribution and acknowledgment using the movie-credit style showed that significantly more contributing individuals and organizations could indeed be identified and documented as compared to the traditional journal citation format. Subsequently, the case study results also demonstrated that if recognition of contribution is an important factor in encouraging and promoting participations in scientific projects, it is equally important to consider and construct a next generation framework for structuring and organizing citation and acknowledgement content, so that a broader range of research efforts can be properly recognized and appreciated.

References

Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science, 35*(2), 180-191. doi:10.1177/0165551508095781

Biagioli, M., & Galison, P. (Eds.) (2003). *Scientific authorship: Credit and intellectual property in science*. New York, NY: Routledge.

Birnholtz, J.P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology, 57*(13), 1758-1770. doi:10.1002/asi.20380

Borgman, C. (2012). What are the attribution and citation of scientific data important? In Paul E. Uhlir (Eds.), *For attribution – Developing data attribution and citation practices and standards: Summary of an international workshop*. doi:10.17226/13564

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing, 28*(2), 151-155. doi:10.1087/20150211

Conley, T. (2003). End credits. In Mario Biagioli, Peter Galison (Eds.), *Scientific authorship: Credit and intellectual property in Science* (pp. 359-368). New York, NY: Routledge.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*, 558-569. doi:10.1002/asi.1097

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855–871. doi:10.1002/asi.10278

Cronin, B., Shaw, D., & La Barre, K. (2004). Visible, less visible, and invisible work: Patterns of collaboration in 20th century chemistry. *Journal of the American Society for Information Science and Technology, 55*(2), 855–871. doi:10.1002/asi.10353

CSE Task Force on Authorship. (2000). Who's the author? Problems with biomedical authorship, and some possible solutions. Report to the Council of Biology Editors (now Council of Science Editors). Retrieved from http://www.councilscienceeditors.org/wp-content/uploads/v23n4p111-1191.pdf

DataCite. (2014). DataCite metadata schema for the publication and citation of research data, Version 3.1. doi:10.5438/0010

Davenport, E. & Cronin, B. (2001). Who dunnit? Metatags and hyperauthorship. *Journal of the American Society for Information Science and Technology, 52*(9), 770-773. doi:10.1002/asi.1123

Edwards, P.N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Flanagin, A., Carey, L.A., Fontanarosa, P.B., Phillips, S.G., Pace, B.P., Lundberg, G.D., et al. (1998). Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. *Journal of the American Medical Association, 280*(3), 222–224. doi:10.1001/jama.280.3.222

Galison, P. (2003). The collective author. In Mario Biagioli & Peter Galison (Eds.), *Scientific authorship: credit and intellectual property in science* (pp. 325-388). New York: Routledge.

Greenberg, J., White, H., Carrier, C., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata, 9*(3-4), 194-212. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/19386380903405090

Greene, M. (2007). The demise of the lone author. *Nature, 450*, 1165. doi:10.1038/4501165a

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

Hou, C.Y. & Mayernik, M. (2016). Formalizing an attribution framework for scientific data/software products and collections. 11th International Digital Curation Conference, Amsterdam, Netherlands.

Hou, C.Y. & Mayernik, M. (2015). The attribution and acknowledgment content framework. Retrieved from http://hdl.handle.net/2142/88845

Osborne, J.W. & Holland, A. (2009). What is authorship, and what should it be? A survey of prominent guidelines for determining authorship in scientific publications. *Practical Assessment, Research & Evaluation., 14*(15). Retrieved from http://www.pareonline.net/pdf/v14n15.pdf

Paneth, N. (1998). Separating authorship responsibility and authorship credit: A proposal for biomedical journals. *American Journal of Public Health, 88*(5), 824-826. doi:10.2105/AJPH.88.5.824

Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. *Eos Transactions, AGU, 91*(34). doi:10.1029/2010EO340001

Parsons, M.A., & Fox, P.A. (2013). Is data publication the right metaphor? *Data Science Journal, 12*, WDS32–WDS46. doi:10.2481/dsj.WDS-042

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails: A proposal to make contributors accountable. *Journal of the American Medical Association, 278*, 579-85. doi:10.1001/jama.1997.03550070071041

Steen, R.G. (2013). Authorship: To be or not to be? *European Science Editing, 39*(1). Retrieved from http://www.ease.org.uk/sites/default/files/esefeb13_essay_rgrantsteen.pdf

Tarnow, E. (1999). The authorship list in science: Junior physicists; perceptions of who appears and why. *Science and Engineering Ethics, 5*(1): 73-88. Retrieved from http://www.onlineethics.org/CMS/2963/resessays/authorship.aspx

Tarnow, E. (2002). Coauthorship in physics. *Science and engineering ethics, 8*(2), 175-190. doi:10.1007/s11948-002-0017-2

Tenopir, C., Allard S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE, 6*(6), e21101. doi:10.1371/journal.pone.0021101

Tilbury, F. (2007). 'Piggy in the middle': The liminality of the contract researcher in funded 'collaborative' research. *Sociological Research Online, 12*(6). doi:10.5153/sro.1644

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics, 6*(4), 700–711. doi:10.1016/j.joi.2012.07.008

Welkos, R. (1998). Giving credit where it's due. Los Angeles Times. Retrieved from http://articles.latimes.com/1998/may/11/news/mn-48618