IJDC | *General Article*

Formalizing an Attribution Framework for Scientific Data/Software Products and Collections

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

Matthew S. Mayernik
NCAR Library
National Center for Atmospheric Research
University Corporation for Atmospheric Research

Abstract

As scientific research and development become more collaborative, the diversity of skills and expertise involved in producing scientific data are expanding as well. Since recognition of contribution has significant academic and professional impact for participants in scientific projects, it is important to integrate attribution and acknowledgement of scientific contributions into the research and data lifecycle. However, defining and clarifying contributions and the relationship of specific individuals and organizations can be challenging, especially when balancing the needs and interests of diverse partners. Designing an implementation method for attributing scientific contributions within complex projects that can allow ease of use and integration with existing documentation formats is another crucial consideration.

To provide a versatile mechanism for organizing, documenting, and storing contributions to different types of scientific projects and their related products, an attribution and acknowledgement matrix and XML schema have been created as part of the Attribution and Acknowledgement Content Framework (AACF). Leveraging the taxonomies of contribution roles and types that have been developed and published previously, the authors consolidated 16 contribution types that could be considered and used when accrediting team member's contributions. Using these contribution types, specific information regarding the contributing organizations and individuals can be documented using the AACF.

This paper provides the background and motivations for creating the current version of the AACF Matrix and Schema, followed by demonstrations of the process and the results of using the Matrix and the Schema to record the contribution information of different sample datasets. The paper concludes by highlighting the key feedback and features to be examined in order to improve the next revisions of the Matrix and the Schema.

Received 20 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Chung-Yi Hou, 1132 Corte Riviera, Camarillo, CA 93010. Email: hou@ucar.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

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.v11i2.404

Introduction

As research topics become more complex and interdisciplinary, the expertise and skill sets required from project team members are also expanding. For instance, in 'Two Librarians, an Archivist, and 13,000 Images: Collaborating to Build a Digital Collection,' Hunter, Legg, and Oehlerts (2010) shared their lessons learned when different professions with distinct skills collaborated together to contribute to the success of a complex project. The project experience demonstrated that the team members were able to tap "into the talents, skills, knowledge, experience, and professional cultures of [their] colleagues." During collaborations, the "convergence across traditional professional boundaries" was critical to ensure the outreach, usability, accessibility, and preservation of data collections. Similarly, in Borgman's (2009) 'The Digital Future is Now: A Call to Action for the Humanities,' she also suggested that collaborations, "when effective, produce new knowledge that is greater than the sum of what the participating individuals could accomplish alone." Her call to "seek out complementary partners" suggests that a fuller range of expertise and cross disciplinary learning is applicable to not only digital humanities, but also all other domains of studies.

While collaborations involving diverse skills and expertise should be encouraged, managing a large, complex project could be challenging. In particular, a significant challenge is distinguishing the contributions of specific people and organizations within collaborative projects. For scientific research, recognition of the different contributing roles and responsibilities could have particular impacts. In terms of academic or professional achievements, receiving official citations or credits for projects could affect career advancement and employment opportunity. Equally important, socially and culturally, acknowledging team members' efforts could help in boosting team morale and esteem, and therefore, potentially enhance accountability and increase work efficiency.

In order to motivate further collaborations among individuals and organizations who can contribute their diverse resources and capabilities for future scientific development, it is crucial to investigate and resolve the issue of how to organize, document, preserve, and disseminate the different contributing roles and responsibilities of a collaborative project, so that proper attributions could be made. A necessary step in building robust mechanisms for collecting attribution information and acknowledgement statements is to formalize appropriate procedures and tools. As a component of a potential solution, the authors have created the Attribution and Acknowledgement Content Framework (AACF), which includes a matrix worksheet and an XML schema that could be used to organize, document, and preserve the variety of contributing roles and responsibilities involved with a scientific dataset/project.

Background and Rationale

The development of the AACF was initiated as a way to compile and organize information about contributors of a large climate-related dataset called the Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40km Reanalysis dataset. The CFDDA was ingested into the Research Data Archive of the National

Center for Atmospheric Research (NCAR) in the summer of 2014 (Hou, Dattore and Peng, 2014). The dataset consists of 183,960 files and is nearly 27TB in volume. In working with the CFDDA creators during the dataset archiving process, it was clear that a traditional citation that listed a small handful of principal investigators did not represent the large number of people and organizations who contributed to the development of the dataset. The development of the AACF was thus motivated by the desire to better account for these numerous and diverse contributions (Hou, Betancourt and Mayernik, 2015; Hou and Mayernik, 2016).

The challenges of effectively attributing responsibility and credit within large scientific projects are well studied (Cronin, 2001; Biagioli and Galison, 2003; Bosnjak and Marusic, 2012). In addition, as discussed further below, a number of schemes have been proposed for categorizing the different types of contributors to scholarly projects. Implementations of these contributor typologies have been limited, though many journals now require authorship and/or contributorship descriptions, particularly in the biomedical domain (Osborne and Holland, 2009). Interest is growing, however, in having authorship and/or contributorship information encoded in more structured and machine-readable form. This would enable more automated compilation of this information across multiple projects and products, as well as enable larger-scale analysis of contributorship in the same way that structured citation information enables traditional bibliometric analyses.

The AACF's goal is to promote a structured representation of contributor information, enabling users to build on existing contributor taxonomies. Collecting contributor information will continue to be an ongoing challenge, but providing additional guidance on how to structure such information might illustrate the benefits more clearly.

Method

When considering the method for constructing an attribution framework, it is critical to ensure that the framework can be adapted for different types of datasets/projects. It is also crucial that the components of the framework are flexible and easy to use, so that the barrier to adoption is kept as low as possible. Focusing on versatility and usability as the main features, the AACF is constructed using a matrix and an XML schema.

The matrix in the two-dimensional format was selected because it allows clear visual representation of the relationship between two different categories. In addition, when structured with multiple rows and columns as grids, the matrix can enable quick comparisons between the categories. Further, there are currently several matrices available that have already demonstrated the effectiveness of summarizing and reporting information through matrix format. Examples of such matrices include the Data Stewardship Maturity Matrix by Peng, Privette, Kearns, Ritchey and Ansari (2015), the NASA Earth Science Data Preservation Content Specification (Ramapriyan and Moses, 2012), and the U.S. Geological Survey Guidelines for the Preservation of Digital Scientific Data (USGS, 2014). As a result, the AACF matrix was created as a two-dimensional grid based on the formats and styles from the referenced matrices.

Although the matrix is a convenient way to provide an overview of the contribution areas and the associated contributing roles for each area, it could be difficult to include substantial details in the matrix, especially for complex projects. An XML schema provides a conduit to implementing the matrix. Many systems (for data, publications,

 doi:10.2218/ijdc.v11i2.404

etc.) already use XML-based metadata for various purposes. Thus, having the AACF instantiated in an XML schema gives a ready-made structure that can be added to existing XML-based systems. Consequently, the authors created the XML schema to facilitate scalable documentation of extensive contributing areas and roles. In order to construct the AACF schema, it was important to consider the levels of description that should be implemented. A balance needed to be achieved between simplicity for schema usage and granularity for record content. By using the format of a data dictionary to model the schema structure, we finalized the following components for the schema: element name, element definition, data type (e.g. string, date/time), format constraint (e.g. ISO 639.2 language code, ISO 8601 date and time format), applicable sub-elements, element obligations (Mandatory, Recommended, or Optional), repeatability/cardinality, applicable attributes, and applicable controlled vocabularies (e.g. Library of Congress Subject Heading). The AACF schema is built according to this data dictionary. Additionally, the current version of the schema is created as a descriptive schema for contribution information only, so that the schema can be integrated into other schemas without causing duplication of information.

In building the AACF matrix and schema, prior works in the construction of schemes for documenting contribution types, roles, and responsibilities were reviewed. In particular, we reviewed the following five different resources:

- Rennie, Yank and Emanuel (1997);
- Paneth (1998);
- Davenport and Cronin (2001);
- DataCite (2014); and
- Brand et al., (2015).

These resources illustrate the multi-decade effort to formalize attribution and contribution information. The most recent of these efforts, by Brand et al., has been formalized into the Contributor Roles Taxonomy (CRediT), which is being implemented in a number of research communities (Singh Chawla, 2015). We were able to consolidate these existing taxonomies and proposals to form a set of 16 contributing types. The 16 contributing types and their associated definitions can be found in Appendix 1. Consolidating these previous efforts and leveraging them within our project allowed us to complement ongoing efforts and promote standardization, as well as to see how these existing contributor type taxonomies overlap. By complementing the AACF matrix and schema with the use of a standard set of contribution types, we aimed to encourage a consistent representation of the information within the AACF matrix and schema.

Results

The current version of the AACF matrix and schema can be accessed online[1]. Although intended to be used as a worksheet for the AACF schema, the AACF matrix could be used independently. The AACF matrix was designed to serve two main purposes: 1) to guide the dataset/project teams' discussions and definitions regarding the contributing areas and the related roles and responsibilities, and 2) to organize and document the

[1] See: http://hdl.handle.net/2142/88845

discussion results. The matrix was constructed using contributing areas as the horizontal axis and contributing types as the vertical axis. Each dataset/project team can select and define the contributing areas that were relevant and specific to their dataset/project. These contributing areas may be based on the stages found in a data lifecycle model, or may be defined as relevant for a particular project/resource.

For the contributing types, the dataset/project could use any of the 16 pre-defined types shown in Appendix 1, or specify other types as needed. For example, the contributor roles defined by CRediT could be used by projects implementing an AACF approach to track contributors.

The matrix also includes an instructions page (shown in Figure 1), a brief background page to describe the dataset/project (shown in Figure 2), and appendices that provide examples of the contributing areas and types that could be used or defined. Using the NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40km Reanalysis dataset as an example, Figure 3 shows the matrix summarizing the contributing types for all five contributing areas defined for the CFDDA dataset (individual and organization names are omitted from the figure for length). The details of the contributing organizations and individuals are documented using the AACF schema.

Attribution and Acknowledgment Content Framework (AACF)
- Matrix Template

Template Version: Revision 1.0 2015-05-11

Instructions:

1) Review the entire file before using the template.
2) Fill out the project/dataset description and the associated contact information on page 2 of this template.
3) Determine the key contributing areas that the related attribution and acknowledge content will be recorded.
4) Record these contributing areas and their associated definitions in Row 1 and 2 of the template starting on page 3.
   a) Additional areas can be accommodated by adding columns to the existing template and continuing on multiple pages.
   b) Helpful Resources:
      i) Data Curation Profiles Toolkit (http://datacurationprofiles.org/overview) could be used to help with the exploration and organization of the contributing areas.
      ii) Examples of contributing areas and their associated definitions can be found under Appendix 1.
5) Review the contributing types/roles and the related definitions.
   a) Helpful Resource:
      i) Details of the contributing types/roles and the related definitions can be found under Appendix 2.
6) For each contributing area, identifications of the related contributing roles can be made by placing a check mark (✓) within the corresponding cells and/or by making any additional notes, as needed.
7) If additional contributing roles need to be defined, add extra rows at the bottom of the matrix and provide the associated definitions.
8) Once the matrix is completed, use the Attribution and Acknowledgment Content Framework XML Schema to create the record for documenting the related individuals and organizations for each contributing area and the identified contributing roles.

Note:

- It is recommended that the matrix to be reviewed and completed as a collaborative process involving multiple representing members of the project/dataset team.
- The matrix should be completed by individual(s) who is(are) able to obtain and provide fair representations of the contributing areas and the related contributing roles.

Citation for this template:

Hou, C.-Y & Mayernik, M. 2015: The Attribution and Acknowledgment Content Framework (AACF) - Matrix Template. Version: 1.0-2015-05-11. Date Accessed: mm/dd/yyyy

Figure 1. AACF matrix instructions page.

 doi:10.2218/ijdc.v11i2.404

Data Attribution and Acknowledgement Content Matrix for <project/dataset> as of YYYY-MM-DD

Revision: #v#

Project/Dataset Title	
Project Homepage or Dataset Landing Page	
Project/Dataset Description	
Project/Dataset Contact (Name; E-mail; Affiliation)	
Data Attribution and Acknowledgement Content Matrix Contact (Name; E-mail; Affiliation)	
Matrix Update History	

Figure 2. AACF matrix project description page.

Data Attribution and Acknowledgement Content Matrix for CFDDA as of 2015-06-30

Revision: 1v0

Contributing Area	Scientific Research Background	Input Files	Software	Data Post Processing	Final Dataset
Contributing Area Definition	The individual or organization who was responsible in establishing the scientific foundation and knowledge base for the production of the dataset.	The individual or organization who was responsible in providing and setting up the initial input conditions for the production of the dataset.	The individual or organization who was responsible in providing, setting up, and maintaining the analysis environment for the production of the dataset.	The individual or organization who was responsible in analyzing, reviewing, and synthesizing the components needed for the production of the dataset.	The individual or organization who was responsible in performing verification and quality control as well as producing the final deliverable version of the the dataset.
Funding and Sponsorship		- 3 organizations.	- 9 organizations and 3 individuals.		- 1 organization and 1 individual.
Project Administration and Management					- 1 organization and 4 individuals.
Software			- 5 organizations and 29 individuals.	- 3 organizations.	- 1 organization and 2 individuals.
Investigation and Raw Data Production		- 3 organizations and 20 individuals.			
Data curation and stewardship		- 4 organizations and 2 individuals.	- 5 organizations and 1 individuals.	- 2 organizations and 14 individuals.	- 2 organizations and 7 individuals.
Final Product	- 52 individuals (as the published authors).				- 1 organization and 9 individuals.

Figure 3. Sample of the AACF matrix for the CFDDA dataset (summary).

Designed as the complementary part to the matrix, the AACF schema should be used to document and store the contribution areas and their respective definitions along with the details of the individuals and organizations who have been identified through the use of the matrix. However, the AACF schema could also be used by itself to create directly the attribution and acknowledgement content records. The schema is structured mainly as a wrapper with several sub-elements. The schema allows one contributing area and its definition to be described per wrapper. Within each wrapper, individuals and

organizations related to each contributing types could be organized and described based on the 16 pre-defined types or by using a self-defined sub-element. Appendix 2 shows parts of the schema and demonstrates a sample of a portion of the attribution and acknowledgement XML record created for the CFDDA dataset using the AACF schema.

Discussion

Since the first demonstration of the matrix and the XML schema of the AACF was based on a climate reanalysis dataset, we are continuing to explore how well the AACF could be applied to additional products that might also result from scientific collaborations. For instance, compared with static datasets, which typically have specific timelines for when individuals and organizations are able to contribute to the datasets and their overarching projects, dynamic datasets evolve continuously on a variety of time scales. There are also various ways in which the datasets can be constructed dynamically. Examples of the different types of dynamic datasets include: 1) on-going projects, such as cruises, that add newly collected measurements to existing datasets at predefined time intervals; 2) regrouping selected data points from an overall dataset to form a specific sub-dataset; and 3) updating an existing dataset with new instrument calibrations to provide an updated version of the dataset. Depending on the data/project type, a dynamic dataset could develop over a significant length of time, and as a result, involve several generations of individuals and organizations.

Similarly, software or simulation models could be shared with and improved by diverse project teams and individuals. In other words, for software that is developed both within a structured team and through collaborations with open communities, ascertaining the revisions made by specific individuals might need particular attention to the mechanisms that could allow and support version tracking and control. This also makes capturing the contribution areas and the related roles for each version of the software and simulation models challenging.

In addition to the CFDDA dataset, which is a static dataset type, the authors have also worked with the National River Flow Archive (NRFA) in the United Kingdom (UK) and Unidata from the University Corporation for Atmospheric Research (UCAR) to evaluate further the applicability of the AACF to dynamic datasets and software respectively. The UK NRFA represents the dynamic dataset use case where datasets from regional hydrometric measuring authorities are submitted to the NRFA on a regular basis in order for the NRFA to generate dataset collections. As for the UCAR Unidata, the production of its NetCDF software is enabled through the combination of dedicated personnel from the responsible organizations as well as any interested members from the software's user communities. As a result, it is a case study that embodies both styles of software development for a software that is used by scientific communities.

When compared to the process and the results of using the AACF matrix and schema to document the contribution information of the CFDDA dataset, both the UK NRFA and the UCAR Unidata NetCDF use cases confirmed that determining the appropriate contribution areas to record the relevant contributing roles and responsibilities could indeed be more challenging for the dynamic than the static data/project types. In other words, since dynamic datasets and projects are constantly evolving, there might not always be a distinct boundary between the different phases of the project or the data's lifecycle. Likewise, it could be difficult to identify definitively

 doi:10.2218/ijdc.v11i2.404

the roles and the responsibilities of the individuals and organizations that could be considered to have provided 'significant' or 'valid' contributions. Furthermore, the contributions to the projects and their associated products could often be either direct or indirect.

For example, in the case of the UK NRFA, the NRFA is responsible for providing data collections to the communities that it supports. However, the data collections are built through the initial data and metadata made available by the individual primary regional hydrometric measuring authorities. Additionally, there are specific roles and responsibilities within the NRFA that ensure the management and the long-term preservation of the final data collections. Meanwhile, the practices of the NRFA are also affected by scientific journals. Major journals are an important stakeholder in the overall science domain and research environment, and can encourage the production of the data collections. Consequently, when creating the UK NRFA's contribution information record, it was important to clarify and correlate the direct contributions to the data collections made by the primary regional hydrometric measuring authorities and the NRFA, as well as the indirect contributions made by the published journals.

Similarly, for the UCAR Unidata use case, the NetCDF software has been developed over a few decades. During this time, NetCDF went through several revisions, and while there were key individuals and organizations that have made direct contributions to the development of NetCDF, many individuals have provided other types of indirect contributions, such as developing additional, complementary tools for NetCDF, reporting bugs, recommending fixes, and providing suggestions. As a result, when recording the contribution areas and roles for the UCAR Unidata use case, it was crucial to indicate the support that the greater community was able to provide to the NetCDF software, in addition to the efforts made by the immediate project developers and funders.

Based on the preliminary assessment, the flexibility of the AACF matrix and schema enables positive indications that the AACF is capable of describing the attribution and acknowledgement content for these two additional project types. For the dynamic NRFA dataset, the AACF matrix and schema can first be used to describe the contributing areas and the related roles for the overall dataset collection. For each sub-dataset with additional, specific contributions, the AACF matrix and schema can then be used again to record the contribution details associated with the particular dataset of the collection. Likewise, for the Unidata software, the AACF matrix and schema can be used to document the contributions made to each identified version. Figure 4 and Figure 5 show respectively the matrices summarizing the contributing areas and types for the UK NRFA and the UCAR Unidata NetCDF use cases (like Figure 3, individual and organization names are again omitted from the figure for length). Finalizing the evaluations with both the UK NRFA and the UCAR Unidata will provide valuable insights regarding the updates that the AACF might need.

Data Attribution and Acknowledgement Content Matrix for UK NRFA as of 2015-08-06

Revision: 1v0

Contributing Area	Monitoring and Network Design	Data Sensing and Recording	Data Validation and Archival	Data Synthesis and Analysis	Initial data and associated metadata provider
Contributing Area Definition	Network of Gauge stations allows researchers to discern climate-driven responses from other, more direct, anthropogenic influences (e.g. abstractions, effluent discharges, reservoirs).	Assessment of the overall quality, including data completeness and homogeneity, and availability of hydrometric information from the gauging network.	Quality control and long-term archiving of hydrometric data.	Provide assessment of the data collected.	Responsible for network installation and upkeep, data collection, data processing, and initial data validation.
Published Journals	- 3 articles and 3 individual authors.			- 10 articles and 16 individual authors.	
Collaborative Programs		- 1 program.		- 1 program.	
Software			- 3 types of software.		
Standardization Institutions		- 3 institutions.			
Primary Regional Hydrometric Measuring Authorities					- 3 organizations.

Figure 4. Sample of the AACF matrix for the UK NRFA use case (summary).

Data Attribution and Acknowledgement Content Matrix for UCAR Unidata NetCDF as of 2015-09-01

Revision: 1v0

Contributing Area	Unidata NetCDF	NASA CDF data model	C Version of CDF-like interfacel	Components of the netCDF package	netCDF-3
Contributing Area Definition	The current version of the NetCDF	Predecessor to the NetCDF data model.	Inspiration for the original NetCDF C interface.	Tools, scripts, documentations, and libraries that constituted the netCDF.	Additional tools, scripts, documentations, and libraries that constituted netCDF-3.
Developers	- 1 organization and 3 individuals.	- 2 individuals.		- 9 individuals.	- 2 individuals.
Funders	- 2 organizations.	- 1 organization.		- 2 organizations.	- 1 organization.
Project Leaders	- 1 organization and 1 individual.				
Hosting Institutions	- 1 organization.				
Collaborators	- 1 organization.				
Distributors	- 1 organization.				
Designers			- 1 individual.		

Figure 5. Sample of the AACF matrix for the UCAR Unidata NetCDF use case (summary).

As the AACF matrix and schema develop further to support the organization, documentation, and storage of the attribution and acknowledgement information for scientific data products and related products, it is also important to consider whether the

 doi:10.2218/ijdc.v11i2.404

AACF can be adapted for other disciplines, such as social science and the humanities. Particularly, the AACF matrix and schema's contributing areas are defined based on the major phases or milestones of a project. In order for the AACF to achieve optimal applicability for all types of research projects, the structures of social science and humanities research projects, and the ways in which individuals and organizations interact and contribute to these projects to produce their results, should also be studied and be integrated into the AACF accordingly.

Finally, as versions of the AACF matrix and schema are finalized, the authors seek to investigate viable methods to implement the AACF for wide distribution. For example, the PaperBadger project[2], a collaboration project among several contributing organizations, has begun building a system that can keep track of individuals who have contributed to a paper, identify their roles based on the CRediT taxonomy as discussed by Brand, Allen, Altman, Hlava and Scott (2015), and assign accordingly the digital badges for the identified contributing roles. By sharing its resources via GitHub[3], the PaperBadger project is openly inviting and involving communities in its development. Consequently, the PaperBadger increases its visibility and possibility to be adopted by a wide range of users. However, the badges currently would only be issued to identify authors of a published paper. It is unclear whether the other project team members who are not part of the author list might receive a badge. Since the AACF aims to be as inclusive as possible when documenting the contribution roles, the authors could consider the PaperBadger's implementation model and evaluate further how such a system could be augmented, and whether other practical alternatives also exist.

Conclusion

Providing appropriate recognition of contributions made to scientific projects and the resulting products, including datasets and software, is an important part of the scientific research process. The attribution and acknowledgement given to the participating individuals and the organizations could not only impact professional reputation and opportunities, but also help in building and encouraging further collaborative behaviour and relationships. Additionally, as scientific projects become more complex and multi-disciplinary, it is also crucial to identify and record the knowledge and skillsets involved, so that the expertise required to produce, manage, and preserve scientific development and output could be understood accordingly.

As part of the potential solutions to assist in the organization, documentation, preservation, and dissemination of the different contributing roles and responsibilities of collaborative projects, the Attribution and Acknowledgement Content Framework (AACF) was developed. The current version of the AACF[4] consists of a matrix worksheet and an XML schema. The matrix and the schema are designed to be versatile and easily incorporated into existing documentation formats, so that the matrix and the schema could be used to record the contribution information of different scientific project/data types.

The AACF was applied to three distinct data types: a static, climate reanalysis dataset (the CFDDA dataset from NCAR); a set of dynamic hydrometric datasets (datasets from the UK NRFA); and an extant software that is continuing to evolve (NetCDF from UCAR Unidata). These three cases enabled an evaluation of the features

2 PaperBadger: https://badges.mozillascience.org/
3 PaperBadger GitHub: https://github.com/mozillascience/PaperBadger
4 Matrix and XML Schema of AACF: http://hdl.handle.net/2142/88845

of the matrix and the schema. As a result of the evaluation process, the AACF demonstrated that it was capable of adapting to different scientific project/data types by allowing and supporting the contributing areas that were unique to each project/data type. The AACF could also include the various roles and responsibilities that were associated with the specific project/data types.

However, in order to achieve optimal effectiveness, the AACF should be applicable to other disciplines beyond the science domain. Many efforts are ongoing across institutions to facilitate the attribution and acknowledge of the contribution information from collaborative projects. As the AACF continues its development, it is vital to explore how the AACF could be enhanced, and ultimately, be integrated with these emerging best practices and mechanisms for attribution and acknowledgement. Recognizing properly the contributions from individuals and organization will help in promoting and supporting collaborative environment in all areas of research.

Acknowledgements

The authors gratefully acknowledge Ethan Davis and Russ Rew of the University Corporation for Atmospheric Research and Harry Dixon of the UK National River Flow Archive for their contributions to this project.

References

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

Borgman, C.L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly, 2009*(3.4). Retrieved from http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Bosnjak, L. & Marusic, A. (2012). Prescribed practices of authorship: Review of codes of ethics from professional bodies and journal guidelines across disciplines. *Scientometrics, 93*(3), 751–763. doi:10.1007/s11192-012-0773-y

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

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*(7), 558-569. doi:10.1002/asi.1097

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

Hou, C.-Y., Betancourt, T., & Mayernik, M. (2015). *Crediting a climate model dataset like a movie? - A case study in data attribution*. Poster session presented at the 10th International Digital Curation Conference, London, UK. Retrieved from http://www.dcc.ac.uk/webfm_send/1805

Hou, C.-Y., Dattore, R.E., & Peng, G. (2014). *Discovering new global climate patterns: Curating a 21-year high temporal (hourly) and spatial (40km) resolution reanalysis dataset*. Poster session presented at the National Center for Atmospheric Research (NCAR), Boulder, CO. Retrieved from https://agu.confex.com/agu/fm14/webprogram/Paper6503.html

Hou, C.-Y. & Mayernik, M. (2016). Recognizing the diversity of contributions: A case study for framing attribution and acknowledgement for scientific data. *International Journal of Digital Curation, 11*(1), 33-52. doi:10.2218/ijdc.v11i1.357

Hunter, N. C., Legg, K., & Oehlerts, B. (2010). Two librarians, an archivist, and 13,000 images: Collaborating to build a digital collection. *Library Quarterly, 80*(1), 81-103. doi:10.1086/648464

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.

Peng, G., Privette, J.L., Kearns, E.J., Ritchey, N.A., & Ansari, S. (2015). A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal, 13*, 231-253. doi:10.2481/dsj.14-049

Ramapriyan, H.K., & Moses, J.F. (2012). NASA Earth science data preservation content specification. Retrieved from https://earthdata.nasa.gov/files/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf

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.

Singh Chawla, D. (2015). Digital badges aim to clear up politics of authorship. *Nature, 526*, 145–146. http://doi.org/10.1038/526145a

USGS. (2014). USGS Guidelines for the preservation of digital scientific data. Retrieved from http://www.digitalpreservation.gov/ndsa/working_groups/documents/USGS_Guidelines_for_the_Preservation_of_Digital_Scientific_Data_Final.pdf

Appendix 1

Table 1. Contributing types pre-defined for the AACF matrix and schema.

Contributing Type Name	Contributing Type Definition
Conceptualization	• Initial conceptualization of the research hypothesis (Paneth, 1998) • Ideas; formulation or evolution of overarching research goals and aims (Brand, Allen, Altman, Hlava and Scott, 2015)
Funding and Sponsorship	• Obtaining funding or material support (Davenport and Cronin, 2001) • Provision and acquisition of the financial support for the project that enabled the creation, management, publication, and maintenance of the results
Supervision	• Entity that is responsible for providing statement of public responsibility (Davenport and Cronin, 2001) • Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team (Brand, Allen, Altman, Hlava and Scott, 2015) and the supervision of other contributing members, such as co-authors
Project Administration and Management	• Coordination of communication among all investigators (Rennie, Yank, and Emanuel, 1997) • Communicating with journal editor (Davenport and Cronin, 2001) • Management and coordination responsibility for the research activity planning and execution (Brand, Allen, Altman, Hlava and Scott, 2015)
Methodology	• Design of the data documentation forms • Development of the study design (Paneth, 1998) • Development or design of methodology; creation of models (Brand, Allen, Altman, Hlava and Scott, 2015)
Writing – Original Draft	• Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation) (Brand, Allen, Altman, Hlava and Scott, 2015) • Including contributing to segments of the original draft

Writing – Review and Editing	• Providing comments on the original as well as all the subsequent drafts • Revision of subsequent drafts from the original draft (Rennie, Yank, and Emanuel, 1997) • Reviewing and critiquing drafts (Paneth, 1998) / Reviewing/proofing manuscript (Davenport and Cronin, 2001) • Provision of final approval of version to be published (Davenport and Cronin, 2001) • Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages (Brand, Allen, Altman, Hlava and Scott, 2015)
Software	• Providing special technical assistance (Davenport and Cronin, 2001) • Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components (Brand, Allen, Altman, Hlava and Scott, 2015)
Investigation and Raw Data Production	• Conducting literature search and analyzing/interpreting relevant literature (Paneth, 1998) • Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection (Brand, Allen, Altman, Hlava and Scott, 2015), extraction, and entry
Formal Analysis	• Data analysis and interpretation of the analyses (Paneth, 1998) • Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data (Brand, Allen, Altman, Hlava and Scott, 2015)
Visualization	• Data analysis (including production of graphs and figures) (Paneth, 1998) • Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation (Brand, Allen, Altman, Hlava and Scott, 2015)
Validation	• Design and execution of validation methods • Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs (Brand, Allen, Altman, Hlava and Scott, 2015)

Data curation and stewardship	• Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse (Brand, Allen, Altman, Hlava and Scott, 2015)
Resources	• Provision of specific material or tools that supported the project, such as data, laboratory samples, raw materials, instrumentation/equipment, computing resources, or other analysis tools
Final Product	• Responsible parties of the final product
Other	• The following activities might also be applicable depending on the project type: ◦ Recruiting human subjects ◦ Mobilizing co-authors ◦ Signing copyright transfer statement

Appendix 2

```
<?xml version="1.0" encoding="UTF-8"?>

<xs:schema xmlns:xs=http://www.w3.org/2001/XMLSchema elementFormDefault="qualified">

   <!-- Attribution and Acknowledgement Content Schema, Rev 1, 2015-05-26 -->
   <!-- XML Schema for recording contributing areas, the associated roles based on contributing categories, and
the relationships among the roles for a scientific project and its resulting products -->
   <!-- The following are the 5 resources referenced in the schema -->
   <!--   1: 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. -->
   <!--   2: Paneth, N. (1998). Separating authorship responsibility and authorship credit: A proposal for
biomedical journals. American Journal of Public Health, 88(5), 824-826. -->
   <!--   3: 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 -->
   <!--   4: DataCite. (2014). DataCite metadata schema for the publication and citation of research data,
Version 3.1. doi:10.5438/0010 -->
   <!--   5: Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: attribution,
contribution, collaboration, and credit. Learned Publishing, 28(2), 151-155. http://dx.doi.org/10.1087/20150211
-->

   <!-- This element sets up the overall container for the entire schema -->
   <xs:element name="AttributionAndAcknowledgementContentSchema" type="ACCSType"/>
   <xs:complexType name="ACCSType">
      <xs:sequence>
         <!-- This is the section that defines the "ContributingAreaAndRole" container -->
         <xs:element name="ContributingAreaAndRole" type="ContributingType" minOccurs="1"
maxOccurs="1">
            <xs:annotation>
               <xs:documentation>
                  The "ContributingAreaAndRole" element is a container for another main sub-container,
"ContributingArea", and all the related sub-elements.
                  The container and its sub-elements are used to document the defined contributing areas and their
related contributing roles based on specific contributing categories.
                     Data Type: none
                     Format Constraint: none
                     Element Obligation: Recommended
                     Repeatability: Non-repeatable
                     Cardinality: If used, must be used only once.
                     Applicable Attributes: none
               </xs:documentation>
            </xs:annotation>
         </xs:element>
      </xs:sequence>
   </xs:complexType>
   <!-- "ContributingAreaAndRole" section -->
   <!-- This is the section that defines the sub-container and all the sub-elements in the
"ContributingAreaAndRole" container -->
   <xs:complexType name="ContributingType">
      <xs:sequence>
         <xs:element ref="ContributingArea" minOccurs="1" maxOccurs="unbounded">
            <xs:annotation>
               <xs:documentation>
                  "ContributingArea" is the sub-container that is used to document specifically the defined
contributing areas.
                     Data Type: none
                     Format Constraint: none
                     Element Obligation: Mandatory, if the "ContributingAreaAndRole" container is used
                     Repeatability: Repeatable
                     Cardinality: If used, can be used at least once
                     Applicable Attributes: none
               </xs:documentation>
            </xs:annotation>
         </xs:element>
      </xs:sequence>
   </xs:complexType>

   <!-- "ContributingAreaName" and "ContributingAreaDefinition" are two sub-elements within the
```

Figure 6. A section of the AACF XML schema.

```
<?xml version="1.0" encoding="UTF-8"?>

<!-- Purpose: This record is created to demonstrate how to use Attribution and Acknowledgement Content
Schema (AACS) to document, organized, and store the different types of contributions of a scientific project. --
>
<!-- Background: The contributions recorded in this record is based on the contributions types identified for the
'NCAR Global Climate Four-Dimensional Data Assimilation (CFDDA) Hourly 40 km Reanalysis
(http://dx.doi.org/10.5065/D6M32STK) -->
<!-- For further information regarding the background work on the CFDDA dataset, please see the following
poster: http://www.dcc.ac.uk/sites/default/files/documents/IDCC15/175_Creatingaclimatemodel.pdf -->

<!-- Date Created: 20150613 -->
<!-- Created By: Chung-Yi Hou (Graduate School of Library and Information Science, University of Illinois at
Urbana-Champaign) -->

<AttributionAndAcknowledgementContentSchema xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="file:/C:/Users/Sophie/Desktop/Misc/AACS_Rev2.xsd">
  <ContributingAreaAndRole>
    <!-- First of Five Contributing Area: Final Dataset -->
    <ContributingArea>
      <ContributingAreaName language="eng">Final Dataset</ContributingAreaName>
      <ContributingAreaDefinition language="eng">The individual or organization who was responsible in
performing verification and quality control as well as producing the final deliverable version of the the
dataset.</ContributingAreaDefinition>
      <AssociatedContributingRoles>
        <FundingAndSponsorship>
          <ResponsibleParty sourceName="NCAR Global Climate Four-Dimensional Data Assimilation
(CFDDA) Hourly 40 km Reanalysis" sourceLink="http://dx.doi.org/10.5065/D6M32STK" affiliation="Defense
Threat Reduction Agency (DTRA) of U.S. Department of Defense (DoD)" type="project sponsor">
            John R. Hannan
          </ResponsibleParty>
        </FundingAndSponsorship>

        <ProjectAdministrationAndManagement>
          <ResponsibleParty sourceName="NCAR Global Climate Four-Dimensional Data Assimilation
(CFDDA) Hourly 40 km Reanalysis" sourceLink="http://dx.doi.org/10.5065/D6M32STK"
affiliation="Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR),
University Corporation for Atmospheric Research (UCAR)" type="sys. admin.">
            John Exby
```

Figure 7. A sample of an AACF XML record.