PACIFIC NORTHWEST AQUATIC MONITORING PARTNERSHIP

Citing Aquatic Monitoring Data Sets: Best Practice Recommendations for Authoritative Data Citation

Sheryn J. Olson, Katie A. Barnas, Margaret R. Williams, Christopher Wheaton, Michael J. Banach, Jennifer M. Bayer

August 2019

Suggested citation:

Olson, SJ, KA Barnas, MR Williams, C Wheaton, MJ Banach, JM Bayer. 2019. Citing Aquatic Monitoring Data Sets: Best Practice Recommendations for Authoritative Data Citation. 33 pages. *available at* https://www.pnamp.org/document/15001

Acknowledgments

The working group volunteered time and insights for seven months and we are grateful for their efforts and participation in this PNAMP project. To promote consensus and cooperation for future efforts this report was reviewed in draft form, and candid comments were solicited at all stages of development. James Duncan generously presented his experiences and processes implementing data attribution for the Forest Ecosystem Monitoring Cooperative. The group benefitted from the experiences, insights and sheer enthusiasm for all things data from Raymond Obuch and Cassandra Ladino, both from the U.S. Geological Survey (USGS). We thank Nancy Leonard of the Northwest Power and Conservation Council, and Russell Scranton of Bonneville Power Administration for productive discussions. We wish to thank the following individuals for their review of the manuscript: Bill Bosch, Yakama Nation; Jay Hesse, the Nez Perce Tribe; and Scott Donahue and Russell Scranton, Bonneville Power Administration. We thank our two reviewers of the final drafts, Cassandra Ladino and Ruth Duerr, PhD. This work was supported by the Pacific Northwest Aquatic Monitoring Partnership (funded by Bonneville Power Administration, project 2004-002-00).

The Pacific Northwest Aquatic Monitoring Partnership is the result of a collaborative effort by many individuals throughout the region. We would like to thank Bonneville Power Administration, National Oceanic and Atmospheric Administration, and the Bureau of Reclamation for their funding contributions. Thank you to all our partners, participants, and collaborators for their continued time and effort in this important endeavor. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Participants in the Work Group

Michael Banach, Pacific States Marine Fisheries Commission (PSMFC); Katie Barnas, Conservation Biology, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration; Jennifer Bayer, USGS-PNAMP; Kasey Bliesner, Oregon Department of Fish and Wildlife (ODFW); Bill Bosch, Yakama Nation; Samuel Cimino, USGS-PNAMP; Nadine Craft, ODFW; Monica Diaz, National Marine Fisheries Service; Scott Donahue, Bonneville Power Administration (BPA); Jay Hesse, Nez Perce Tribe; Cassandra Ladino, USGS; Raymond Obuch, USGS; Lenora Oftedahl, Columbia River Inter-Tribal Fish Commission; Sheryn Olson, USGS-PNAMP; Tom Pansky, BPA; Russell Scranton, BPA; Ryan Santo, Nez Perce Tribe; Rebecca Scully, USGS-PNAMP; Samantha Smith, Nez Perce Tribe; Christopher Wheaton, PSMFC; Mari Williams, Ocean Associates- Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA.

Citing Aquatic Monitoring Data Sets: Best Practice Recommendations for Authoritative Data Citation

Sheryn J. Olson[1], Katie A. Barnas[2], Margaret R. Williams[3], Christopher Wheaton[4], Michael J. Banach[4], Jennifer M. Bayer[1]

Keywords data citation, data attribution, salmonid recovery, data sets, Pacific Northwest, monitoring, salmon, metadata

Abstract

The use of data generated from long term monitoring efforts necessitates accurate authoritative source citations of those data to ensure credit for data collected, and accountability for the data quality to enable repeated retrieval of a given data set. Data sets used in published reports and articles are increasingly being considered objects that are required to be published and cited. Aggregating data into open access databases is becoming common and is the focus of the Coordinated Assessment for Salmon and Steelhead project (CA; https://www.pnamp.org/project/coordinated-assessments-for-salmon-and-steelhead; http://www.streamnet.org/data/coordinated-assessments/) and National Marine Fisheries Service, National Oceanic and Atmospheric Administration Salmon Population Summary (SPS; https://www.webapps.nwfsc.noaa.gov/apex/f?p=261:home:0) among others. Guidelines are needed for citing these long-term dynamic data sets that have many contributors. We explore best practices and provide recommendations for including robust metadata attributes within data sets to enable data publication and citation using the CA and SPS data repositories as case studies. From reviewing the current citations possible from the CA and the SPS we recommend at minimum that natural resource monitoring databases contain: metadata to identify organizations that generated the data; contact persons for each organization that contributes data to an aggregated data set; and that metadata be incorporated into databases to enable auto-generated citations that recognize all contributing organizations with time-stamped versions of the data delivered. Beyond those minimums, additional best practice recommendations include this suite of metadata elements that identify a given data set upon citation or publication: author(s); publication date; description of data; file format(s) of data - e.g. tiles, shapefile sets, images, text files; dates data were collected; locations where data were collected; producers/contributors to the data set version cited; date data set was downloaded; original data repository from which the data were obtained; version identifier to note significant change to a data set; and a persistent identifier that can be used to locate that version of the data.

[1] Pacific Northwest Aquatic Monitoring Partnership, U.S. Geological Survey, Cook, Washington

[2] Conservation Biology, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd E. Seattle, Washington 98112, USA

[3] Ocean Associates, Ocean Associates, Inc. contracted to Northwest Fisheries Science Center, National Marine Fisheries Service NOAA. Portland, Oregon

[4] Pacific States Marine Fisheries Commission, Portland, Oregon

Contents

Figures and Tables

Introduction

Many organizations in the Pacific Northwest collect significant amounts of fish and fish habitat data, at times to inform management and recovery of Endangered Species Act (ESA) listed Pacific salmon. Tribal, state, and federal organizations monitor aquatic resources and contribute to many data repositories, for water quality, physical habitat measures, and fish population parameters, among others[1]. Metadata that describe attributes or characteristics of data, the data acquisition processes, and data contributors can provide a mechanism for the successful reuse of data at multiple scales and help to enable collaboration. Compiling data from many sources and summarizing them in one data set necessitates the creation of data sharing agreements, metadata standards, and consistently implemented human and machine-readable metadata. Here, we consider how to effectively cite portions of a database or aggregated data sets for aquatic monitoring programs. We explore and recommend options for data attribution and citation for two databases that house salmon abundance and productivity trends.

Why start with data citation and attribution? Data citation is but one aspect of data governance and management. Having data governance policies and data management plans in place can best foster collaboration, data accessibility, and re-use of data. Best practices would be to manage aquatic data throughout the Pacific Northwest with consistency among the many federal and state agencies, tribes, universities, partnerships and private entities. Additionally, implementing metadata management is a critical piece of data management, with three key phases: planning; implementing and maintaining operational metadata management; and steps to improve metadata management (Obuch 2018).

We focused on a small section of these larger processes for three reasons: 1) among many of the Pacific Northwest Aquatic Monitoring Partnership (PNAMP) organizations, there was great interest in giving data collectors, stewards, processors, and analysts due credit; 2) some federal agency partners had a need for a mechanism to publish their data along with reports and articles; 3) this was an achievable, feasible project. A comprehensive regional set of data governance policies and a data management system will be a much larger project. Numerous entities have data management systems, and agreement among them to reach a common, regional system will be a long term and gradual process.

This paper arose out of the Coordinated Assessments for Salmon and Steelhead project (CA: https://www.pnamp.org/project/coordinated-assessments-for-salmon-and-steelhead) which brought together regional partners to create the Coordinated Assessments (CA) database for salmon abundance and productivity data. CA participants recognized the need for metadata to easily find and access data, and to easily produce auto-generated citations. We identified recommendations for best practices as a priority when it became clear that all data providers needed to be credited before the aggregation of many organizations' data could be shared and used collaboratively. Data from aquatic research, monitoring, and evaluation in the Pacific Northwest are an asset that deserves protection and preservation (NPCC 2016). However, differing data sharing and access rules apply to the many partners involved in the CA project. For example, beginning in

[1] Water quality data: USGS Waterwatch, State of Washington Dept. of Ecology River and Stream Monitoring; physical habitat metrics: Columbia Habitat Monitoring Program, NOAA CHaMP metrics, Integrated Status and Trends Organization; indicators for ESA listed fish species: http://www.fpc.org/, https://fishandgame.idaho.gov/ifwis/portal/, http://www.odfwrecoverytracker.org/; and fish population parameters: https://fortress.wa.gov/dfw/score/score/, http://www.cbr.washington.edu/dart/overview, http://www.streamnet.org/data/coordinated-assessments/. These are a few examples of data repositories among many.

2013, federal agencies receiving more than $100 million for research and development, such as NOAA's Northwest Fisheries Science Center (NWFSC), were directed to develop plans to make the results of federally funded research freely available to the public within one year of collection, and to better account for and manage data resulting from federally funded scientific research (OSTP 2013). While a number of executive orders since 2009 apply to accessibility and transparency of federal agencies' data (e.g. NWFSC and USGS[2]), data sharing policies are less stringent for contributing state and tribal entities. Information generated from all these efforts is a valuable asset that can inform best practices in science and policy, and well-documented data sets can improve trust, accountability and the ability to reuse data (Altman 2011, ESIP 2019, Ponzio 2004, Starr *et al.* 2015). This is especially critical for data produced by authoritative sources (Ponzio 2004, Appendix D).

Differences in data management mandates among PNAMP partners led the PNAMP Steering Committee and several partners to request that PNAMP facilitate a workgroup to address two goals: 1) recommend best practices for data attribution and citation for consideration by PNAMP partner organizations, and 2) consider current data attribution and citation processes for two specific data repositories commonly used by partners. The intent was to inform best practices recommendations and to illustrate the uses and applications of data attribution and data citation.

Subsequent to the release of these recommendations – called Phase One – the Data Citation and Attribution Working Group may proceed with further efforts in future phases if there is interest among partners to continue the project. In Phase Two we would implement our optimal recommendations for two existing databases to enable thorough data citation and to allow data set retrieval. The PNAMP project would include evaluating successes and lessons learned from the implementation in Phase Two. Many participants in the original working group anticipated a Phase Three to present the recommendations and results from the project as a journal publication to involve a larger audience of monitoring practitioners than PNAMP partners in the Pacific Northwest.

We began by reviewing current best practice recommendations with a focus on the region's need for accurate data acknowledgment and citation for PNAMP partners. We reviewed information from the global community of earth science monitoring practitioners, policy implementers, natural resource managers, and researchers. The working group sought expertise from groups and individuals who had implemented metadata management programs to enable thorough data citation and re-use of data sets: the Forest Ecosystem Monitoring Cooperative, US Department of the Interior's Wildland Fire Management, and members of the USGS Community for Data Integration (see Appendices B, C).

We assessed as case studies two databases commonly used by partners for what could currently be cited upon download from the repositories: NOAA Northwest Fisheries Science Center's (NWFSC) salmon population summary database (SPS: https://www.webapps.nwfsc.noaa.gov/apex/f?p=261:home:0), and the Coordinated Assessments database (https://www.streamnet.org/data/coordinated-assessments/) for fish population indicators. We then compared those citations to citations of dynamic scientific and monitoring data sets from the global community that use recommended best practices (ESIP 2019).

[2] An open data policy was mandated in 2009 for federal agencies that release publications that present data (OMB M-10-06). Executive Order "Making Open and Machine Readable the New Default for Government Information" (OMB M-13-13 2013), required such data to be accessible, persistent, well documented with metadata, and citable. A framework to do so was outlined in OMB Memo-13-13.

We developed recommendations with the participation of interested representatives from tribes, states, and federal agencies that perform fish monitoring in the Pacific Northwest. Future efforts can include implementation of pilot programs to provide optimal data attribution of data repositories (as Phase 2), with potential application to programs beyond aquatic fisheries and fish habitat programs (Phase 3).

As we explore best practices to enable data citation, it is helpful to clarify the terms "attribution" and "citation". Though data citation is not equivalent to data attribution (Appendix D), data sets must have been accurately assigned metadata elements for citation to be possible, and for those data sets to be located and reused in that same state. The term "data attribution" can be used with different meanings: 1) to provide credit for, that is to attribute or acknowledge an entity or person that was responsible for some portion of data collection or handling, or 2) metadata characteristics (attributes) attached to a data set as pre-defined fields that describe a data set, enabling citation of a data set as an independent object similar to a journal article. For the purposes of this paper, we use the broader sense of data attribution – data attributes as metadata characteristics or metadata elements that are assigned to a given data set. Credit or acknowledgement for data collection and handling is only one of many metadata elements that can be assigned to a data set.

Rationale

This effort addresses the concern surrounding appropriate use of data sets that result from several organizations collaborating to conduct research and monitoring. Accurate documentation of provenance, or tracking, of a resulting data set and associated metadata becomes critically important to provide credit and accountability. Future authors or analysts may selectively use a partial data set or combine disparate data sets inappropriately. This type of use has been labeled as data parasitism (Longo and Drazen 2016). Further, reusing data without any citation or acknowledgement is essentially plagiarism (Duke and Porter 2013). Such inappropriate use of data can distort or eliminate provenance. Of greater concern is that the use of a data set cannot be replicated lacking metadata, and upon reuse, results may be misinterpreted, possibly contradicting or inaccurately reflecting the original authors' results (Longo and Drazen 2016). Though it is impossible to prevent purposeful misuse of data, implementing data attribution best practices can reduce the risk of accidental misuse and increase trust in the quality and provenance of the data.

Providing metadata as a component of data stewardship is essential to be able to publish data sets with accurate citations (Borgman 2011). Data citations are critical for several reasons (Altman 2011):

- Appropriate attribution – Data citations enable appropriate legal and scholarly attribution for the cited work.
- Persistence of data and metadata – Citations refer to, and allow management of, objects that are persistent.
- Access – Citations are used to facilitate short and long term access to the object, by humans and by machine clients.
- Discovery – Citations are used to locate instances of the data set and as part of the process of discovering derivative, parent, and related works.
- Provenance – Citations are used to associate published claims with evidence supporting them, and to verify that the evidence has not been altered. (Altman 2011)

Data used to inform and implement policy thus requires persistence, provenance, credibility and accountability. Evaluation of long-term data in population biology can lead to a better understanding of populations' status and trends, provide comparisons of multiple watersheds, and illustrate multi-scalar effects of habitat restoration (action effectiveness) or degradation, examine food webs and their resilience, and reduce uncertainty when combining data from many populations of exploited species. For example, data from the CA database represents nearly 400 salmon and steelhead populations, and the ability to produce reliable meta-analyses or generate trusted reports from aggregated data sets will better inform future policy decisions and assist the region's co-managers to implement management strategies identified as best practices. Agencies such as the NWFSC and StreamNet (www.streamnet.org) provide authoritative data, i.e., officially recognized data that can be certified and provided by an authoritative source (Ponzio 2004, Appendix D), and authoritative data need to meet standards of quality and ensure veracity and trust in the data.

Best Practices for Data Citation

Data citation standards lagged behind scientific literature citation standards, though since 1997 task forces and international organizations have been developing and urging standards for data attribution, citation, discoverability, validation and documentation to improve transparency, integrity and reuse (FGDC 1998, Starr et al. 2015, ESIP Data Preservation and Stewardship Committee 2019). Inaccessible data raise concerns about reproducibility, for example, false positives may be reported as true positives (Colquhoun 2014, Rekdal 2014). In 2013, referencing data sets with bibliographic information was referred to as "The relatively new practice" and "first principles" were recommended for data citation (Soche 2013). As recently as 2015, many agencies and global consortia were in the process of determining how and what will work for data citation and attribution, especially how to best implement standardized attribution and metadata elements for multi-source data across disciplines and organizations (Starr et al. 2015, Sprague et al. 2017). This working group seeks to contribute to the discussion around development of best practices, and to encourage PNAMP partners to adopt universally recognized best practices. Thorough data citation has yet to be instituted as a common practice among most PNAMP partners, and the recognition of the critical need for acknowledging data contributors led to this paper. We first explore the state of best practices that are globally recognized and current standards for metadata to enable thorough data citation (Obuch et al. 2018, ESIP Data Preservation and Stewardship Committee 2019).

Effective, comprehensive data stewardship improves data accessibility and interoperability, and ensures trust of the data sets and data systems, and of the people participating (Parsons and Fox 2013). Some journals (e.g. Ecology) and federal agencies now require, and many others strongly suggest (e.g. Canadian Journal of Fisheries and Aquatic Sciences) that data sets used in a study are simultaneously published as separate "data objects" that are retrievable and reusable. To be able to cite and publish a data set, that data object must be attributed with persistent identifiers (locators), and a set of metadata identifiers.

Two efforts have established baseline standards for data management and attribution: FAIR data principles and the Joint Declaration of Data Citation Principles. FAIR data principles were developed by a group of organizations belonging to FORCE11 that formed a working group in 2015 to define good data management, and provide guidance for data to be Findable, Accessible, Interoperable and Reusable, the FAIR concept (Martone 2015, Wilkinson, et al. 2016). The Coalition for Publishing Data in the Earth and Space Sciences

(COPDESS.org) had 43 publishing and professional society signatory agencies by 2017, including such organizations as American Association for the Advancement of Science, American Geophysical Union, Dryad data repository, Nature Publishing Group, National Snow and Ice Data Center, Proceedings of the National Academy of Sciences, and the International Council for Science-World Data System (https://www.icsu-wds.org/) with its 110 member organizations as of February 2018. COPDESS is committed to encouraging FAIR data principles and the Joint Declaration of Data Citation Principles, and to work toward improving quality of, and access to data repositories.

The Joint Declaration of Data Citation Principles was endorsed by 115 organizations globally as of September, 2017. (Martone 2014, FORCE11.org). These eight guiding principles recognize that citation practices need to be human understandable and machine-actionable to support reuse of accessible, robust data.

1. Importance: Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. Credit and Attribution: Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. Evidence: In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. Unique Identification: A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. Access: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. Persistence: Unique identifiers, and metadata describing the data, and their disposition, should persist -- even beyond the lifespan of the data they describe.
7. Specificity and Verifiability: Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
8. Interoperability and Flexibility: Data citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data citation practices across communities.

Minimally the following metadata elements should be present in data set descriptions (Starr *et al*. 2015):

- Data set_Identifier — A machine-actionable identifier resolvable on the internet, a persistent landing page, that points to the data set (Janée *et al.* 2009)
- Title — The title of the data set.
- Description — A description of the data set, with more information than the title.
- Creator — The person(s) and/or organizations who generated the data set and are responsible for its integrity.
- Publisher_Contact — The organization and/or contact who published the data set and is responsible for its persistence.
- PublicationDate/ Year/ReleaseDate — ISO 8601 standard dates are preferred (Klyne & Newman, 2002).

Additional recommended metadata elements in data set descriptions for best practices are:

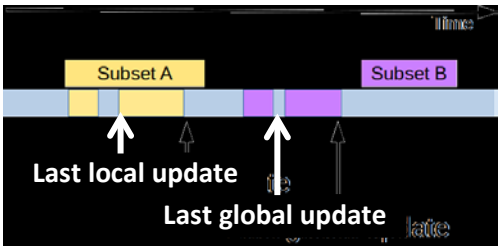
- Creator identifier(s) — ORCiD or other identifier of the individual creator(s) in addition to name.
- License — The license or waiver under which access to the content is provided (preferably a link to standard license/waiver text (e.g. https://creativecommons.org/publicdomain/zero/1.0/).
- Version — The data set version identifier (if applicable)

The challenge of dynamic data sets

Data generated from long term research, monitoring, and evaluation programs are dynamic, benefitting from iterative processing, error correction, refinement of data collection techniques and study designs, and addition of data records from collaborators (Parsons and Fox 2013). Using persistent identifiers such as Digital Object Identifiers (DOIs) and Life Science Identifiers (LSIs) to attribute static data sets has had mixed success (Duerr *et al.* 2011). Dynamic data sets further challenge those who need to use portions thereof (Rauber *et al.* 2016). So, to accurately cite dynamic data for potential reuse requires a suite of data management processes. These include effective, distributed governance of the data system, data preservation, providing accessibility to both humans and machines, and reasonable credit and accountability for data collection, creation, and curation, and crucially, versioning for dynamic data sets (Parsons and Fox 2013). This creates challenges that are both technical and socio-cultural (Uhlir 2012).

The PNAMP working group agreed that the indicators in the CA database needed accurate provenance and that organizational credit and accountability were paramount. Additionally, for optimal data retrieval, citation, and reuse, versioning and multiple time-stamps are necessary, especially for reusing subsets of a dynamic data set. Query stores, which allow a query to be saved and re-executed, are one mechanism to handle dynamic data. This avoids having to store whole subsets of the data that were queried with their time stamp and retriever. Who downloads, what portion of the database they download, and a time-stamp for when they download, can be imported as metadata fields if the structure of the database allows it – the query store would work as its own table with a key link to the data set downloaded. Metadata records of the person and organization that made the query and what data sets were downloaded are the stored query table. Four processes have been proposed as mechanisms to attribute data in dynamic data sets that need to be retrieved as subsets (Rauber *et al.* 2016).

- Prepare the data – version the data with a time-stamp, and the query store



(source: Rauber *et al*. 2016)

- Mark data set downloads at the time of the query as Persistent Identifiable Data sets (PID), UUID preferred with time stamp and version (Duerr *et al.* 2011)
- Resolve PIDs and Retrieve the Data
- After the data infrastructure is modified, verify successful data and query migration

These procedures and recommendations are not comprehensive; there will be additional tasks and procedures identified as more thorough data attribution is implemented.

Data citation of aquatic monitoring data sets in the Pacific Northwest

Fisheries time series data, like most long-term natural resource monitoring efforts, are dynamic and subject to modification, which demands particular attention to provenance and accountability. Population level salmon abundance and productivity data differ from data collected for a study with a defined time frame in that 1) multiple co-managers may contribute to a time series, 2) time series annual values are often recalculated for past years based on new data, and 3) data are often reused or recombined.

One request of the Data Attribution and Citation working group was that for population indicators in the CA database, adding metadata elements begin with sufficient, but not exhaustive levels of attribution. The intent was to simplify both data input and data retrieval. The following list describes policies and technical approaches for data attribution and citations identified by the larger group of all partners conducting monitoring, research and evaluation:

- Data sharing agreements should be established for all online data repositories.
- End User License Agreements (EULAs) should be in place for every instance of data downloads.
- Implementation must be feasible and non-intrusive.
- Automated associations of data sources should be attached to all downloaded data.
- Auto-generation of a citation when data are downloaded.
- Ability to cite consolidated sets of records, combined indicators, and data subsets
    - Inclusion of data source field(s) in aggregate databases from contributing organizations
    - Versioning for additions/changes to data sets
    - Time stamp for version changes
    - Version of data downloaded for use and citation
    - Time stamp for data subset download
    - Addition of persistent identifier, or locator, when publishing and citing a data set

We suggest these items as best practices for aquatic monitoring data in the Pacific Northwest in accordance with others in the earth and life sciences global community who support establishing best practices for data management, preservation, attribution and citation (Appendix B, Borgman 2011, ESIP 2019, Martone 2014, Rauber *et al.* 2016, Starr *et al.* 2015).

Two Case Studies for Aquatic Monitoring Data Repositories: Status Quo

As background we first present an overview of the characteristics of two databases and what is possible to download and cite from those databases as of July 2019. Members of the PNAMP working group who have implemented metadata structures for optimum data citation and retrieval noted that it is crucial to assess the status quo of the database architecture to be able to add metadata necessary for thorough data citation. That includes identifying existing metadata fields and metadata gaps, which would affect resulting schema design.

We explored two data repositories: The Coordinated Assessments for Salmon and Steelhead Populations (CA) database and the Salmon Population Summary (SPS) database, managed by StreamNet and the National Marine Fisheries Service, respectively, that are commonly used by PNAMP partners. We first assessed what could currently be cited when using data sets from these databases and then explored recommendations to implement metadata necessary to enable citation of partial databases, or aggregated data sets.

Case Study 1: Coordinated Assessments for Salmon and Steelhead Populations database

The National Marine Fisheries Service defined parameters to evaluate salmonid populations for recovery, known as viable salmonid population (VSP) parameters, that were intended to characterize the long-term viability of naturally spawning salmonid populations (McElhany *et al*., 2000). Commonly referred to as "population indicators", methods and criteria have been developed by Columbia River basin Technical Recovery Teams (NOAA's Northwest Fisheries Science Center (NWFSC) Conservation Biology Division). Values of these indicators are estimated from sets of metrics and measurements that have been summarized or synthesized, and in many cases produced from multi-step statistical modeling. Data collection and analyses are complex. For example, multiple organizations may have contributed data to inform the development of one indicator, necessitating accurate attribution of these indicators when, for example, these indicators are aggregated for reports across watersheds, over multi-year time periods, or population data are used to develop predictive models.

The CA project developed a Data Exchange Standard (DES: https://www.streamnet.org/coordinated-assessments-des/) to enable capturing VSP data in a central database in a unified data structure to facilitate data sharing and use. The CA DES defines data table structures, field definitions, and data integrity rules for several population-scale indicators. These indicators are: natural origin spawner abundance (NOSA); smolt to adult return rate; recruits per spawner for adult and for juvenile recruits; number of juvenile outmigrants; number of parr; and proportionate natural influence in integrated natural/hatchery populations (Table 1).

Table 1. Viable Salmon Population (VSP) indicator names and descriptions for natural origin fish populations. VSP indicators were developed by technical recovery teams sponsored by National Marine Fisheries Service (McElhany et al. 2000) and are captured in the CA database. The indicators are used to evaluate the status of salmon and steelhead populations. The table is modified from Coordinated Assessments Data Exchange Standard Version 20170701 (Pacific States Marine Fisheries Commission, StreamNet Project DES 2017).

VSP Indicator	Description
Spawner abundance	Number of natural origin fish that actually spawn, not necessarily the number of fish returning to a spawning area.
Presmolt abundance	Number of natural origin juvenile fish in a population. Usually late summer parr but may be any time and stage.
Number of outmigrants	Number of fish passing a defined point as they migrate downstream.
Smolt to adult ratio (percentage)	100 X the point estimate of the number of returning natural origin adults, divided by the point estimate of the number of smolts that produced those returning adults.
Recruits per spawner	Recruit per spawner ratios are specific to the locations and seasons described in each record of data. The number of "recruits" can be defined at any life stage.
Proportionate natural influence	Estimate of the relative selection pressure of the natural environment in an integrated natural/hatchery population.

Each record in the CA represents one indicator for one fish population for one year. Several metadata fields in the database provide the ability to attribute or credit organizations for each data record (Table 2). The *ContactAgency* field captures the main organization that calculated the indicator value in each data record. The *OtherDataSources* field is used to document additional organizations that contributed to creation of the data record, so that they can be credited. The *RefID* field links to a table of references and references can be documents, personal communications, memoranda, etc. These three fields, along with fields to document methodologies, provide information on how the data were created and by whom. In addition, information on who to contact for more information is included with each data record. The *OtherDataSources* field is optional and can contain up to 255 characters for contributing organizations, separated by the pipe symbol (|). Agency names are verified by validation rules, so a record will not load until all agency names included exactly match the standard list of agencies.

Using the available metadata fields, automated routines can be implemented to machine auto-generate citations from aggregated records, eliminating duplicate metadata such as one contact person per many records. Another optimal set of desired metadata elements will include Persistent Identifiers linked to controlled versioning of data sets, keywords, descriptors of the type, and format of the data, and a persistent data repository identifier (or URL). These elements could be easily added to the current tables and could ensure that downloaded and versioned data sets could be accessible and reused as they were in previously used states.

Table 2. This excerpt from the Natural Origin Spawner Abundance (NOSA) Data Exchange Standard (DES) table displays metadata fields describing data used to estimate the NOSA indicator in the Columbia River basin. The table is excerpted and modified (Pacific States Marine Fisheries Commission, StreamNet Project DES 2017).

Field Name	Field Description	Data Type	Codes/Conventions for NOSA Table
	Protocol and method documentation		
ProtMeth URL	URL(s) for published protocols and methods describing the methodology and documenting the derivation of the indicator	***Memo***	Required if ProtMethDocumentation is null. Provide URL(s) to source documentation of methodology. For MonitoringMethods.org provide link to the protocol.
OtherData Sources	The ContactAgency field identifies an organization involved in calculating the values in this record.	Text 150	List all the organizations that provided data used to calculate the values for this record. If more than one, separate the entries with the bar character " | ". This field is for ADDITIONAL organizations.
	Supporting information		
NullRecord	In some years data may not be collected and so indicator values cannot be calculated. For example, high muddy water or wildfires can prevent redd counts that indicator values are based on. This field is used to indicate that indicator values do not exist because the data do not exist to calculate them.	**Text 3**	Normally "No". A value of "Yes" in this field is a positive statement that the data do not exist to calculate the indicator for the population and time period specified. Metric data and age data may still exist when NullRecord = "Yes". missing data are explicitly accounted for
DataStatus	Status of the data in the current record.	**Text 255**	Acceptable values: Draft, Reviewed, Final
Last Updated	Date (and time if desired) the information in this record was last updated.	Date/Time	If an exact date is not known, give an estimate.
Indicator Location	Where this indicator is maintained at the source.	Memo	If online, provide URL(s).
Metric Location	Where the supporting metrics are maintained at the source.	Memo	If online, provide URL(s).
Measure Location	Where the measurements are maintained that were used for these calculations.	Memo	If online, provide URL(s).
Contact PersonFirst	First name of person who is the best contact for questions that may arise about this data record.	**Text 30**	
Contact PersonLast	Last name of person who is the best contact for questions that may arise about this data record.	**Text 30**	
Contact Phone	Phone number of person who is the best contact for questions that may arise about this data record.	**Text 30**	Preferred format is "123-456-7890".
Contact Email	Email address of person who is the best contact for questions that may arise about this data record.	**Text 50**	

Data and metadata are available from the online CA query system as a downloadable spreadsheet. The spreadsheet includes all indicators. It may include all populations, or instead only data for population(s) selected (Figure 1). The record level metadata of each indicator for a fish population by year may be too fine-grained to meet the needs of users of the data who need to aggregate records for purposes such as a five-year status report. A field was added in 2017 to enter data contributors and authors who calculated the estimations for each record. The field is a text string that can include many data collectors and authors from

many organizations. A mechanism to attribute different person-roles or organization-roles is not yet implemented, e.g. organization-person-data collector, organization-person-data steward, and organization-person-data analyzer.

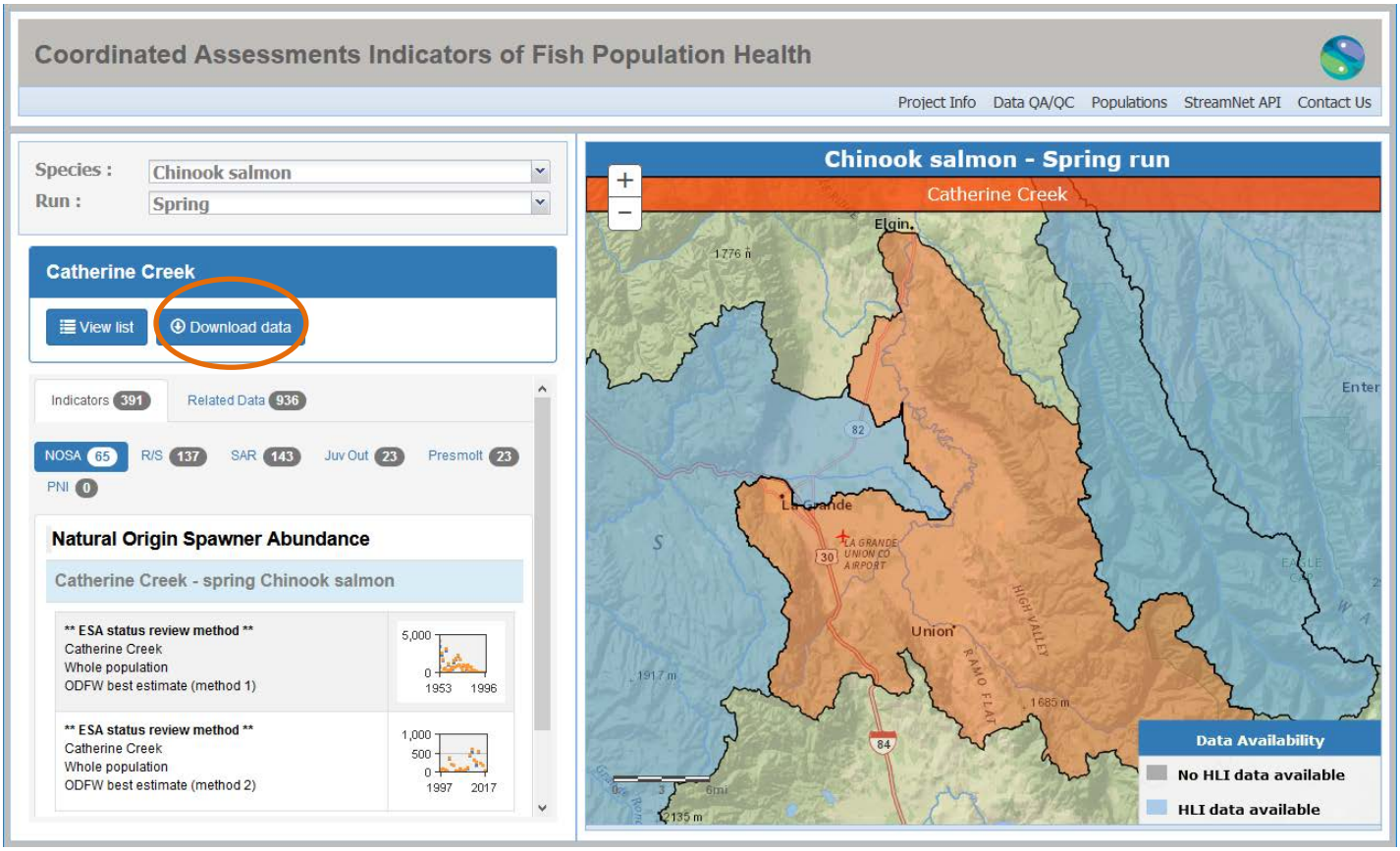


Figure 1. A screen shot of the Coordinated Assessments query system, displaying the Natural Origin Spawner Abundance (NOSA) indicator data available for a single population of fish, Chinook spring-run salmon in Catherine Creek, Northeast Oregon. "Indicators" are highly derived population-scale data characterizing population viability: population size (NOSA, Juveniles Out, and Presmolt); reproductive rates (R/S); survival rates (SAR); and hatchery influence (PNI). "Related Data" are of more limited geographic scope and are generally indexes of abundance summarized from field data. Both types are time series of annually calculated values. The circled "Download data" button provides a spreadsheet of all available data for the selected population(s). (Screen shot from Coordinated Assessments query system: http://cax.streamnet.org/?species=Chinook%20salmon& run=spring&popid=7##). Accessed 2019-07-19. Further information about Coordinated Assessments can be found at https://www.streamnet.org/data/coordinated-assessments/ and https://www.pnamp.org/project/coordinated-assessments-for-salmon-and-steelhead.)

While data are available for download, there is no mechanism to machine-generate a citation from database fields, nor to tag a subset of downloaded data with certain metadata such as a version of the data set, and type of data (e.g. geodatabase, shapefile set, ASCII file, CSV file). However, much of the information is available in the data sets as existing fields, and time-stamps are attached to changes in the database, so metadata information could be extracted to provide complete citations. Minor additions or alterations to the database structure, for example, in the form of relational attribute tables, XML scripts, and user generated

report templates could provide flexible, machine auto-generated citations. Current excel files of data that can be downloaded when the CA database is queried contain many columns, the including metadata fields that could be used to generate a citation. Some of these existing metadata fields, listed below, could potentially be used in a data set citation.

COMMONPOPNAME
WATERBODY
SPAWNINGYEAR
CONTACTAGENCY
INDICATORLOCATION {URL}
CONTACTPERSONFIRST
CONTACTPERSONLAST CONTACTPHONE
CONTACTEMAIL
DATA ENTRY {Person}

A subset of those metadata can currently be used to manually derive a citation such as this example:

Idaho Department of Fish and Game (2018-09-12). Idaho Department of Fish and Game, Oregon Department of Fish and Wildlife, Nez Perce Tribe (NPT), and US Forest Service, data contributors. Dataset version: July 17, 2015 1:15 pm. Recruits per spawner, Snake River Chinook spring run ESU, 2003-2015. Excel spreadsheet. Accessed at Coordinated Assessments 2017-05-02: http://cax.streamnet.org/?species=Chinook%20salmon& run=spring&popid=7##, See Supplemental Table A1 for original data source citation details.

In sum, the many existing fields in the current database enable quite thorough citation of data sets. For example, a machine-generated citation could be attached to queries that includes the person who requested the download; download date/time; location of where the data were collected for records downloaded; data description including indicator(s) name and fish population name; all entities contributing data to the records; database use time/date stamps to show when data were changed; a persistent identifier assignment; and the name of the authoritative, persistent data repository that houses the data set. The reuse of any particular dataset would require a persistent identifier (or locator) to retrieve the data set version as it existed on the citation date.

Case Study 2: NOAA Salmon Population Summary Database, Northwest Fisheries Science Center

The Salmon Population Summary (SPS) database, managed by NOAA's Northwest Fisheries Science Center (NWFSC), was designed to provide access to demographic data compiled for ESA-listed salmonid populations as part of the NWFSC's technical recovery planning efforts. The database contains data on spawning abundance, age structure of wild spawners, fraction of natural spawners that are of wild origin, and the reduction in spawning abundance due to harvest. The data correspond to the populations identified by NOAA's National Marine Fisheries Service (NMFS) Technical Recovery Teams, and are used in part to assess population and evolutionarily significant units (ESU)-level recovery criteria for many listed ESUs, e.g. for five-year status reviews.

Data in SPS are updated once annually, made available for public use, and a new archive is created. All past data archives are provided, and past archives can be queried. Current citations may take the form of a NOAA contact person for each Recovery Domain who manages or synthesizes data for that domain, knows the individual agency contacts, and communicates with them. Sometimes all people are listed but often only the contact name, then the agency or tribes that provided data (Table 3).

Table 3. Example citations that are currently possible from the NOAA's Northwest Fisheries Science Center (NWFSC) database. Citations include fish populations, population indicator names, and years when data were collected. In examples 1 and 2, M. Rowse is the NWFSC lead for the domain (Puget Sound) and T. Johnson is the tribal lead. In example 3 T. Cooney is the NWFSC lead for the Interior Columbia Recovery Domain, and he conferred with co-managers at ODFW, Yakama Nation, WDFW and the Nez Perce Tribe to compile his estimates.

Example	Fish Population	Indicator	Citation and Attribution	Data Years
1	Chum Salmon (Hood Canal Summer-run ESU) - Strait of Juan de Fuca	Spawner Reference	Compiled by M. Rowse from CoManager Five Year Status Review (2005-2013) and data excel spreadsheet provided by T. Johnson, Point No Point Treaty Council. 2015. Annual Estimates of Population Abundance, Age, Hatchery/ Supplementation Contribution, Productivity: Hood Canal and Strait of Juan de Fuca Summer Chum	1971-2013
2	Chum Salmon (Hood Canal Summer-run ESU) - Strait of Juan de Fuca	Spawner Reference	T. Johnson. 2014. T. Johnson, Point No Point Treaty Council	
3	Steelhead (Snake River Basin DPS) - Grande Ronde River Upper Mainstem Summer-run	Spawner Reference	Compiled by T. Cooney, NWFSC from reports and data provided by ODFW, YN Fisheries Division, WDFW, and NPT Fisheries. 2011. Annual Estimates of Population Abundance: Snake River Steelhead DPS -Upper Grande Ronde	2007-2010

For data sets to be reused and cited as their own objects, a given data set must be retrievable in the future. Examples 1 and 3 may provide enough information to retrieve the data set, if one is able to contact T. Johnson, M. Rowse or T. Cooney. To retrieve any particular data set you would need a locator to where that data set resides, which is generally a persistent identifier (such as a Digital Object Identifier or DOI). These citations are manually developed, rather than auto-generated from fields in the SPS database.

Recommendations

An ideal data citation that conforms to global best practices ensures that the data set cited displays enough metadata so that the data set can be reused and has a persistent location that allows retrieval of that data set in the same state as was cited. During Phase One, developing recommendations, the PNAMP Data Attribution and Citation working group focused on these elements, which we address in terms of our two case studies:

- Determine levels of thoroughness for attribution of desired citations to describe data subsets or aggregated data
- Recommend **minimum** levels of data attribution
- Recommend **optimum** levels of data attribution to allow accurate citations and data interoperability: sharing, access, subsequent retrieval, and reuse.

In accordance with these elements, we recommend attributing data sets in existing repositories at a minimum to credit data contributors and provide accountability. A further recommendation is to provide a version of a data set, or a sub-set of data that can be used as a cited data object, at reasonable intervals that reflect database alterations. A third, optimal recommendation is to provide a persistent identifier to retrieve a data set cited in the same condition as it was cited. One initial, feasible recommendation to provide

accessibility with unique identifiers is to assign the CA database a DOI number, and the SPS database a different DOI. Then, when a portion of the database is used, a universal unique identifier (UUID) can be requested for the sub-set of the data that will have the time stamp and description of which portion was used (R. Duerr pers communication, May 2, 2019, Duerr et al. 2011).

Levels of thoroughness for data attribution elements of desired citations

Attribution and the number of metadata elements attached to a data set can be more detailed than the level of detail in a citation. Not all metadata elements need to be displayed in a citation. Citations of a data set can be auto-generated when metadata are included or attached to a database. Machine readable metadata values can be used to build a citation, and are desirable to enable queries of a database to retrieve a given data set. With more metadata fields that describe attribution, citations can have more or less detailed metadata information. Retrieval of a given data set is only possible when a persistent locator or a persistent identifier such as the Digital Object Identifier (DOI) is present as exhibited in the last three citations in Table 4, so even a minimal citation of a dataset needs a unique identifier.

Table 4. Citations with increasing levels of thoroughness. Comments describe how metadata may be used to generate a citation and whether the cited data are retrievable. The first three examples were citations proposed by the Nez Perce Tribe (Hesse and Wheaton 2014), while the last three examples have data set versions and persistent identifiers, or locators.

Level	Citation	Comment
Low	Data utilized in this paper generated by the Nez Perce Tribe, Washington Department of Fish and Wildlife, Idaho Power Company, and US Fish and Wildlife Service; downloaded from the Fish Passage Center (FPC.org - 1/30/2015).	Not auto-generated but could be coded as such. With no version, data may not be retrievable in same state as it was in 1/30/2015.
Adequate	Idaho Department of Fish and Game, Oregon Department of Fish and Wildlife, Nez Perce Tribe (NPT), and US Forest Service. As accessed through www.sps.org July 17, 2015 1:15 pm. See Supplemental Table A1 for original data source citation details.	Fields can be auto-generated. Data may be retrievable from Table A1. Original data source citation details could be included here.
Detailed List of Data Contributors	Apperson and Janssen 2005. Apperson 2004, 2008. Ball et al 1979. Bjornn and Richards 1961. Bjornn et al 1966. Brown 2000. Corley and Welsh 1968. Corley et al 1971, 1973. Elms-Cockrum 1996, 1997, 1998, 1999, 2001. Elms-Cockrum 1995. Hall-Griswold and Cochnauer 1986, 1988a, 1988b. Hassemer 1993. Hauck 1954. Holubetz and Van Vooren 1977. Holubetz et al 1966. Hoss et al 1975, 1976, 1977. Idaho Dept. of Fish and Game Unpublished, 2010, 2011, 2012, 2013. Apperson and Janssen 2006. Leth et al 2000. Lindland et al 1980. Metsker 1957. Nemeth 1988, 1999. Ortmann and Richards 1964. Ortmann et al 1964, 1965, 1981. Pirtle and Simpson 1957. Pollard 1983, 1984, 1985. Pollard et al 1982. Rabe and Nelson 2010, 2011, 2012, 2013, 2014. Rich et al. 1992, 1993, 1994. Richards and Gebhards 1959. Richards and Bjornn 1960. Riley and Elms-Cockrum 1995. Sharr 2003. Holubetz et al 1971. Vogel 2005. Vogel and Nelson 2006. Welsh et al 1966, 1968, 1969. Welsh et al 1972. Welsh et al. 1977. White and Cochnauer 1988.	Fields can be auto-generated. Data location to a persistent repository are not described, data set not retrievable.
Ideal	Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 2007, updated daily. MODIS/Aqua Snow Cover Daily L3 Global 500m Grid V005.3, Oct. 2007- Sep. 2008, Tiles (15,2;16,0;16,1;16,2;17,0;17,1). Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-11-01 at doi:10.1234/xxx.	Can be Auto-generated: Each element is retrievable as a metadata field. DOI ensures same data packet can be retrieved.
Ideal	Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated 2003. CLPX-Ground:ISA snow depth transects and related measurements, Version 2.0, shapefiles. Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at doi:10.1234/xxx.	Can be Auto-generated
Ideal	Zwally, H.J., R. Schutz, C. Bentley, J. Bufton, T. Herring, J. Minster, J. Spinhirne, and R. Thomas. 2003. *GLAS/ICESat L1A Global Altimetry Data V018, 15 October to 18 November 2003*. National Snow and Ice Data Center. data set accessed 2011-07-21 at doi:10.3334/NSIDC/gla01.	Can be Auto-generated

Ideal machine-generated citations are possible when metadata fields have constrained values in user input forms rather than free-form text fields (See Appendix B: Best practices for data citation, this document). Such citations can be machine-generated to produce authors, data subset description, version, type of data, where, and when data were collected, data source, date accessed, and a persistent identifier that is a locator (DOI or UUID) to the data set.

Minimal recommendations for citing aquatic data sets

Monitoring programs in the Pacific Northwest collaborate extensively with many organizations designing monitoring efforts and studies and collecting data. Later, other organizations may summarize and synthesize the data from these multiple sources. Additional organizations may analyze those summaries and report results. It is critical to credit all of the organizations that contributed and handled the data, both for acknowledgement and accountability.

Various agencies, organizations and tribes in the Columbia River basin have different needs for use of the data sets. They therefore require varying degrees of how comprehensive a data citation need be.

If the goal is to acknowledge the contributing organizations that provided data, then fewer metadata attributes are necessary. Simply a description of the indicator, one or more organizations that handled data, date range of data collection and a location can be sufficient to provide acknowledgement of contribution. However, if the goal of a data citation is to provide enough information to retrieve that data set in the same state and from the same time it was cited, then at the minimum, citations need to include the bolded items below, some of which would be cross-referenced in a database, and databases would need corresponding fields to allow for machine-generated citations.

- Organization
  - **Organization name**
  - **Organization contact information**
  - Organization contact person (cross referenced to person affiliation)
  - Organization role(s): Data provider-contact person/steward; Data analyst; Data validator (governance); Data contributor
- Person
  - Person affiliation (cross referenced to Organization)
  - Person roles: study designer/principle investigator; data manager or steward; Data validator (QC); data analyst; report writer; data provider; data contributor.
- Data description
  - **Persistent Identifier** (DOI or other location identifier to data set)
  - Identify metadata location (e.g. links to MonitoringResources.org)
  - Metadata header records: What-description; **Type of data** - description and format, e.g. text or comma separated values file, shapefile set, images, tiles; **Geospatial description**, **date initiated, date completed**; how collected – links to procedural metadata, authorization (link to organization authorizer; person authorizer)
- Data provenance
  - **Time stamps: downloaded as a set**, database changes
  - **Versioning** – Predetermined amount of change to assign version with person who changed, or added to database.

Coordinated Assessments Database Viable Salmonid Population Indicators

The CA database has many fields that could be used when citing a data set, and when data are downloaded, there is a time stamp added to the downloaded database. However, no persistent identifier is attached to the extremely dynamic database. It is not possible to limit queries by a time frame when data were collected. After downloading a Microsoft Excel file, it is possible to use a portion of the records. Each record has

additional metadata links, time frames when data was collected, and organizations that contributed data. But no persistent identifiers or locators to that specific data download are included.

Because personnel may change frequently, the working group recommends acknowledging the organizations as raw data sources. Minimum recommendations for inclusion as attributes were organization name of data contributor and contact person for that organization. Included as metadata with all downloads will be an End User License Agreement. Some in the working group suggested that certain metadata fields be included in the CA database DES and contain specific values:

Organization roles – Study designer, Data collection, Data validator, Data analyzer, Report writer
Person roles – Principle investigator, Data analyst, Data QA/QC validator(s), Contact steward

Attributes specific to salmon population data repositories can include elements that currently exist as database fields or can be added as fields: population name, parameters (indicators), locations, and time stamps accompanying database changes and downloads. An example of a minimal citation from the CA query system may be:

Data used in this paper generated by the Nez Perce Tribe, Washington Department of Fish and Wildlife, Idaho Power Company, and US Fish and Wildlife Service; downloaded from the Fish Passage Center (FPC.org - 1/30/2015).

Here, the specific set of data used is not findable again, accessible, or reusable as cited, though it does accurately credit organizations who contributed data. The first implementation for adding metadata to the CA database was to acknowledge data contributors. This was discussed and approved at the CA workshop meeting on May 11, 2017. Pursuant to the agreement, StreamNet staff added the fields to the CA Data Exchange Standard released July, 2017 to capture and acknowledge organizations that contribute data. Persistent identifiers are generally considered to be a minimum recommendation for data citation and publication. To include a persistent identifier locator such as a DOI, it may be necessary for an author using a portion of the database to assign one when the data are published with a specific publication, because persistent identifiers are not assigned with data downloads or attached to subsets of the database.

Salmon Population Summary database, National Marine Fisheries Service

Viable Salmonid Population indicators in the National Marine Fisheries Service SPS database are currently obtained from the CA database. The SPS database has a five-year cycle corresponding to the five-year reviews of listed populations, and database versions are finalized and made static at these five-year intervals. At that point, the database can be assigned a DOI. There has been interest in using the database in the interim, and the interim uses would necessitate a different mechanism for citing the data sets. Minimum desired data attribution would allow citation of data by year and data type, with the description of the indicator. Acknowledgment of data contributors to the indicator and the data repository is another minimum recommendation. For purposes of reusing a data set, a contact person or organization is also required. For example, a single time series may have multiple references, each representing one or more years of data. When possible, National Marine Fisheries Service personnel cite publicly available reports that explain how data were collected and summarized. In turn, those citations become integral to the data sets downloaded by the consumers. The system of attribution is much improved on past versions, though agencies are more likely

to be attributed (e.g. ODFW, NPT) rather than the numerous biologists at each agency that contributed. A proposed example format of such a citation is:

Chum Salmon (Hood Canal Summer-run ESU) - Strait of Juan de Fuca	Spawner Reference Data years 2001-2013	Compiled by M. Rowse from CoManager Five Year Status Review (2005-2013) and data excel spreadsheet provided by T. Johnson, Point No Point Treaty Council. 2015. Annual Estimates of Population Abundance, Age, Hatchery/Supplementation Contribution, Productivity: Hood Canal and Strait of Juan de Fuca Summer Chum

This citation requires one to contact M. Rowse and possibly T. Johnson to obtain the same data set that they cited. It also assumes that Rowse and Johnson have kept the data set in the same condition as when it was last used. When data have become distributed, as it was to M. Rowse in this example, the data set may not have been preserved in the same state as it was when cited.

Optimal recommendations for citing aquatic data sets

Elements to include as metadata to enhance a given data set's retrievability and reusability are: Person-roles such as principle investigator, data steward, data QA/QC validator, data analyzer; data subset description; download time stamp; type of data (e.g. text or comma separated values file, photos, shapefile set, images, tiles); organizational roles such as funders, study designers, data analyzer, report writer; geographical description with coordinate system, datum, and attributes; persistent identifier; URL of source repository; and version of data set accessed (see Appendix B: Data Citation Standards). An entire database such as the CA or SPS is a data set, and a best practice recommendation is to assign a Digital Object Identifier to each that would provide access to the entire live data set and can be attached as a metadata field upon download. For partial use of a dataset, a universally unique identifier (UUID) can be chosen by a user, when the user uses a partial dataset and metadata attached to that data set can be a complete description of which portion and a time stamp or version number when downloaded.

Coordinated Assessments Database: Viable Salmonid Population indicators

Roles for people and for organizations involved, persistent identifiers, or relational links to metadata can be added as the working group and reviewers deem necessary. Many of these attributes exist in the current CA database, though they are attached to the data at the record level. For instance, fields exist for download date, date last modified, validation date, references, contact person, and others. Because the data sets can be used for legal or regulatory reasons, and it may be necessary to aggregate several data sets in a given report, there is a critical need for citations that describe aggregated or subsetted data. An ideal citation would include more than the above bolded attributes, for example:

Data User(s) or Citers, Date published, Data Collectors, Data set name, dates & time span when data were collected, Data type, Location of data (geophysical description), Organization validated data, version, Persistent Unique ID or Location of data set {URL}, Access date

Ideal citations in Table 4 are derived from extremely dynamic, online databases that change daily. When the CA is queried, there could be several authors and organizations. Date that the data were used and the persistent identifier, the DOI, must be added manually, unless metadata software is used to manage data sets upon use. One example of an optimal, potential data set citation from the CA database could be:

Idaho Department of Fish and Game, Oregon Department of Fish and Wildlife. 2018. Data generated by Nez Perce Tribe (NPT), and US Forest Service. Major Population Group: Lower Snake River – Tucannon River Spring Chinook Salmon, Natural Origin Spawner Abundance. 1954-2016. V2.0. Nez Perce Tribe maintained at www.nptfisheries.org. Excel file. DOI 10.xx.xxxx. Accessed at http://cax.streamnet.org/?species=Chinook%20salmon&run=spring&popid=15##, July 17, 2017.

This is one possible citation for illustration purposes. Implementation of these examples can be machine generated and easily implemented using much of the existing database structures, as illustrated by the Forest Ecosystem Monitoring Cooperative's implementation (Appendix C).

Salmon Population Summary Database, National Marine Fisheries Service

Using the above citations from the National Marine Fisheries Service current examples (Table 3), optimal citations could take the following form shown in the second example here, by consolidating and adding information into the citation. Note the sample DOI reference and access date. This is an example of a potential citation for illustration purposes only.

Current	Chum Salmon (Hood Canal Summer-run ESU) - Strait of Juan de Fuca	Spawner Reference 2001-2013	Compiled by M. Rowse from CoManager Five Year Status Review (2005-2013) and data excel spreadsheet provided by T. Johnson, Point No Point Treaty Council. 2015. Annual Estimates of Population Abundance, Age, Hatchery/Supplementation Contribution, Productivity: Hood Canal and Strait of Juan de Fuca Summer Chum

Optimal

Rowse, M, compiler. 2017. Johnson, T, Point No Point Treaty Council, Apr 2015, data provider. Type: excel spreadsheet. And Five Year Status Review (2005-2013). Annual Estimates of Population Abundance, Age, Hatchery/Supplementation Contribution, Productivity: Hood Canal and Strait of Juan de Fuca Summer Chum ESU – Strait of Juan de Fuca. 2001-2013. V. 3.2. Data set accessed 2017-07-07 at doi:10.5555/NOAAnmfs/sps05.

Note the addition of 1) date cited, 2) data type or format description (e.g. spreadsheet, text file, shapefile set, images), 3) data set version, 4) date accessed, 5) DOI. The example persistent identifier (Digital Object Identifier) points to a persistent online repository where the data are stored and accessible as a packet that can be downloaded in the future in the same state as it was on 2017-07-07. Because the SPS database is finalized as a static product every five years, that static product can be assigned a DOI. If there is a possibility that an interim download is used, that time-stamped (or partial) download can be assigned a UUID (Duerr *et al.* 2011).

Discussion

Phase One of this PNAMP task – Recommendations – concludes with the release of this white paper containing best practices recommendations for data attribution and citation for aquatic monitoring practitioners. We recommended minimum and optimum metadata elements to enable data citations for two organizations' data repositories: The Coordinated Assessments Exchange for salmon and steelhead and the National Marine Fisheries Service salmon population summary indicators. Included in the best practices

recommendations are metadata elements necessary to enable thorough data citations of authoritative data. Phase One sets the groundwork for Phase Two.

Phase Two – Implementation – would involve providing resources and the participation of several groups to implement these further recommendations:

- Practitioners of research, monitoring and evaluation, policy developers and implementers (managers) need to continue to provide more examples of ideal citations for aggregated population data when using data sets for meta-analyses, management, legal challenges, or for reports (Table 4).
- Resources need to be identified and allocated to work with database architects, database managers and developers, and end-users to educate them to use data attribution.
- Data stewards and data architects need to identify the metadata gaps in the current data repositories for data published via the Coordinated Assessments Exchange, and NOAA's Salmon Population Summary database.
- Data architects and programmers need to develop database schemata, data dictionaries, and report generators to retrieve data sets with attached metadata fields. Included with data downloads, metadata fields are necessary to machine-generate a data set citation. These standardized fields could also be used for report queries.
- Develop a data governance framework to address quality of metadata, including citation enforcement, and build on metadata element completeness, initially validating Not Null metadata element entries to determine that all metadata elements contain valid values.

Following more complete implementation of metadata fields for data citation, a useful exercise would be to assess costs and benefits of implementation at different levels of thoroughness. Upon approval and agreement among all interested entities, StreamNet staff could lead implementation for the CA database with participation from data architects, stewards, and managers, and from organizations providing data. The PNAMP workgroup would welcome the opportunity to be involved in subsequent phases, pending recommendation of such tasks by the PNAMP steering committee and partners. This PNAMP workgroup, the Data Attribution and Citation Working Group, has gained experience about metadata standards for data citation, and learned lessons for working with each other that will be valuable for facilitating further implementation of metadata standards and practices among diverse agencies, tribes and interested stakeholders.

This task group could later consider mechanisms to standardize and operationalize the implementation or continued use of best practices for metadata necessary to enable data citation. That process would entail implementation of a data governance plan, and ongoing processes. Data governance plans will vary by the agency, organization or tribe that chooses to use them and involve implementation using various approaches, such as a non-invasive or a traditional approach (Seiner 2016). Resources for learning about data governance, creation of data governance plans, and general guidelines for data governance are attached (Appendix A).

With long-term monitoring efforts producing considerable amounts of valuable data, data management is also a necessary portion of the data life cycle to assure data quality during data collection, subsequent data handling and preservation of data (Chatfield *et al.* 2011). Data management is a larger issue than data citation and includes such processes as managing metadata and addressing ways to best cite data sets. Ensuring data are thoroughly and accurately documented with metadata can promote the ability to

accurately cite data. In the Columbia basin, there is a project in progress to implement data management among tribes, and as the project matures, the concepts in this paper can contribute to dissemination of data. The Inter-Tribal Monitoring Data Project has the purpose of assisting the Columbia River Inter-Tribal Fish Commission (CRITFC) and its member tribes in the timely and accurate capture, storage, processing, and dissemination of data for management of anadromous fish and their habitats (CRITFC 2008-2019, Roe 2019). The CRITFC project can serve as a model for other aquatic monitoring organizations in the Pacific Northwest as they implement data management plans.

References Cited

Altman, M. 2011. Data Citation in The Dataverse Network Presentation Prepared for the Board on Research Data and Information "Developing Data Attribution and Citation Practices and Standards. An International Symposium and Workshop" August 22-23, 2011

Borgman, C. 2011. Why are the attribution and citation of scientific data important? Presentation at Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop. August 22-23, 2011. Accessed on the World Wide Web 28 Sept 2016 *available at* http://sites.nationalacademies.org/PGA/brdi/PGA_064366

Chatfield, T., Selbach, R. February, 2011. Data Management for Data Stewards. Data Management Training Workshop. Bureau of Land Management (BLM).

Colquhoun D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science 1(3):140216 *available at* https://royalsocietypublishing.org/doi/full/10.1098/rsos.140216

Columbia River Inter-Tribal Fish Commission (CRITFC). 2008-2019. BPA Project #2008-507-00 - CRITFC Inter-Tribal Monitoring Data. https://www.cbfish.org/Project.mvc/Display/2008-507-00. Accessed 2019-07-12

Coordinated Assessments Exchange for Salmon and Steelhead. 2017. www.streamnet.org, managed by StreamNet: Fish Data for the Northwest. *available at* http://cax.streamnet.org/?species=Chinook%20salmon&run=spring&popid=7##. Accessed 2017-04-08

Duerr, R.E., R.R. Downs, C. Tilmes, B. Barkstrom, W.C. Lenhardt, J. Glassy, L.E. Bermudez, P. Slaughter. 2011. On the utility of identification schemes for digital earth science data: an assessment and recommendations. Earth Sci Inform (2011) 4:139–160.

Duke, C.S. and J.H. Porter. 2013. The ethics of data sharing and reuse in biology. BioScience. 63(6):483-489.

ESIP Data Preservation and Stewardship Committee. 2019. *Data Citation Guidelines for Earth Science Data. Ver. 2*. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816. accessed Jul 2, 2019

Federal Geographic Data Committee. (FDGC) 1998. "Content Standard for Digital Geospatial Metadata" Version 2 - 1998. (FGDC-STD-001 June 1998) *available at* https://www.fgdc.gov/csdgmgraphical/index.html accessed Apr 10, 2017

Hesse, J. and C. Wheaton. 2015. Giving Credit Where Credit Is Due: Ethics and Standards for Using Online Data sets. Presentation. The American Fisheries Society Annual Meeting. August 16-20, 2015. Portland, Oregon *available at* https://afs.confex.com/afs/2015/webprogram/Paper20953.html

Janée G., J. Kunze, J. Starr. 2009. Identifiers made easy. Available at http://ezid.cdlib.org/.

Klyne, G. and C. Newman 2002. RFC3339: date and time on the internet: timestamps. *available at* http://www.ietf.org/rfc/rfc3339.txt.

Longo, D.L. and J.M. Drazen. 2016. Data Sharing. Editorial, The New England Journal of Medicine. 374:3

Hesse, J. and C. Wheaton. 2015. Giving Credit Where Credit Is Due: Ethics and Standards for Using Online Data sets. Presentation. The American Fisheries Society Annual Meeting. August 16-20, 2015. Portland, Oregon *available at* NOAA Salmon Population Summary Database. 12 Feb 2013. NOAA Northwest Fisheries Science Center. Scientific Data Management. Seattle, Washington 98112. USA. Ver. 2.0 available online at https://www.webapps.nwfsc.noaa.gov/apex/f?p=261:HOME:::::: accessed 5 Apr 2017.

Northwest Power and Conservation Council (NWPCC). May 2016. Annual Report. www.nwcouncil.org 851 SW 6th Ave., Suite 1100, Portland OR 97204. Figures 6, p.16 and 10, p.21. *available at* https://www.nwcouncil.org/media/7150247/2016-4.pdf accessed Apr 14, 2017

Martone M. (*ed*.) 2014. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].

Martone, M.E. 2015. FORCE11: Building the Future for Research Communications and e-Scholarship. Bioscience 65:635.

McElhany, P., M.H. Ruckelshaus, M.J. Ford, T.C. Wainwright, and E.P. Bjorkstedt. 2000. Viable salmonid populations and the recovery of evolutionarily significant units. U.S. Dept. Commerce, NOAA Tech. Memo. NMFS-NWFSC-42,156 p

Obuch, R.C., Carlino, Jennifer, Zhang, Lin, Blythe, Jonathan, Dietrich, Chris, Hawkinson, Christine, 2018, Department of the Interior metadata implementation guide—Framework for developing the metadata component for data resource management: U.S. Geological Survey Techniques and Methods, book 16, chap. A1, 14 p., https://doi.org/10.3133/tm16A1.

Office of Management and Budget (OMB) Memorandum M-10-06. Open Government Directive. December 8, 2009. *available at* https://www.osehra.org/sites/default/files/us-omb_open-government-directive_m10-06.pdf

Office of Management and Budget (OMB) Memorandum M-13-13. May 9, 2013. Open Data Policy – Managing Information as an Asset. *available at* https://project-open-data.cio.gov/policy-memo/

Office of Science and Technology Policy (OSTP) Memorandum. Feb 2, 2013. Increasing Access to the Results of Federally Funded Scientific Research. *available at* https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf accessed Apr 14, 2017.

Pacific States Marine Fisheries Commission, StreamNet Project. 2017. Coordinated Assessments Data Exchange Standard (DES) Development Team. CAX DES. Ver 20170701. 61p. *available at* http://www.streamnet.org/coordinated-assessments-des/

Ponzio, F.J. 2004. Authoritative Data Source (ADS) Framework and ADS Maturity Model. Proceedings of the Ninth International conference on Information Quality (ICIQ-04). 346-357 *available at http://ssm-vm030.mit.edu/ICIQ/Documents/IQ%20Conference%202004/Papers/AuthoritativeDataSourceFramework.pdf*

Rekdal, O.B. 2014. Academic urban legends. Social Studies of Science 44(4):638‑654 https://doi.org/10.1177/0306312714535679

Rauber, A., A. Asmi, D. vanUytvanck, S Pröll. 2016. Identification of Reproducible Subsets for Data Citation, Sharing and Reuse. Research Data Alliance Working Group for Data Citation. Online: https://www.rd-alliance.org/system/files/documents/TCDL-RDA-Guidelines_160411.pdf

Roe, C. 2019. CRITFC Inter-Tribal Monitoring Data. Presentation, Portland, Oregon.

Seiner, R. 2016. Comparing Approaches to Data Governance. The Data Administration Newsletter. DATAVERSITY Education, LLC *available at* http://tdan.com/comparing-approaches-to-data-governance/20386. Accessed 20171003

Starr *et al.* 2015. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Comput. Sci. 1:e1; doi 10.7717/peerj-cs.1 *available at* TGDCSPhttps://peerj.com/articles/cs-1.pdf

Sprague, L.A., G.P. Oelsner, D.M. Argue. 2017. Challenges with secondary use of multi-source water-quality data in the United States. Water Research 110:252-261.

Task Group on Data Citation Standards and Practice (TGDCSP). *editor* Y.M. Socha. 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 12, pp.CIDCR1–CIDCR7. https://doi.org/10.2481/dsj.OSOM13-043

Uhlir, P.F., Rapporteur. 2012 . For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop National Research Council of the National Academies. The National Academies Press, Washington, D.C. *available at* http://www.nap.edu/catalog.php?record_id=13564

Wheaton, C. 2017. Coordinated Assessments Update. Presentation. Pacific Northwest Aquatic Monitoring Partnership (PNAMP) Steering Committee Meeting. April 6, 2017. Portland, Oregon.

Wilkinson, M. D. *et al*. (52 others). 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Appendix A. Resources

Key documents and notes from the working group are posted on PNAMP.org in the Data Attribution Project, the Pacific Northwest Monitoring Partnership's website.

Learning about Data Management: The Earth Science Information Partners (ESIP) has a training website with several modules. Especially pertinent is **Lesson 8: Data Citation**.

Metadata in Plain Language: While data attribution isn't all of metadata, this list is a good start and all attribution needs to include many of these considerations (What, Who, Why these data, How, where collected?). This list considers the importance of describing geospatial elements of data – such as datum, extent, point-line-polygon, coordinates, etc.

NOAA NCEI Metadata Training: The National Centers for Environmental Information (NCEI) provides metadata training to help data providers and data managers accomplish the goal of providing discovery-level, access-level, and understanding-level metadata for their geospatial and environmental data. NCEI's metadata training focus is on the development of ISO 19115-2 and related standards in accordance with NOAA's Data Documentation Directive: https://www.ncddc.noaa.gov/metadata-standards/metadata-training/

Importance of data dictionary **What is a Data Dictionary?**

A data dictionary is a set of information describing what type of data was collected within a database, its format, structure, and how the data set is used. In many respects, a data dictionary can be thought of as the rules in which all the data within your system need to abide by. If all of your systems are producing data that follow the same rules - you achieve semantic interoperability. Example USGS Data Dictionaries:

- EarthExplorer USGS Landsat Data Dictionary
- Data Dictionary for USGS Water-Use Data
- U.S. Geological Survey Open-File Report 03-001: Data Dictionary for Surficial Sediment Data from the Gulf of Maine, Georges Bank, and Vicinity: A GIS Compilation

Recommended Reading

Ball, A. & M. Duke. 2011. How to Cite Data sets and Link to Publications. DCC How-to Guides. Edinburgh: Digital Curation Centre.

Costello, M.J. 2009. Motivating Online Publication of Data. BioScience 59(5):418-427.

DataCite - International Data Citation. 2011. DataCite Metadata Schema for the Publication and Citation of Research Data [PDF]. Version 2.2.

Data Observation Network for Earth: 2016. DataONE education modules.

Earth Science and Information Partners (ESIP) Federation Interagency Data Stewardship/Citations/ Provider Guidelines. And ESIP Data Preservation and Stewardship Committee. 2019. *Data Citation Guidelines for Earth Science Data. Ver. 2*. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816.

Hakala, J. 2010. Persistent identifiers - an overview. Standards in Metadata and Interoperability. Technology Watch Report. Online.

Schimel, D. 2017. Open Data. Frontiers in Ecology. 15(4):175.

US Geological Survey. 2017. Data Citation. USGS Data Management website

US Geological Survey. 2018. Guidance on Documenting Revisions to USGS Scientific Digital Data Releases. USGS Fundamental Science Practices website

Pennington, Deana, University of New Mexico, and Shawn Bowers, LTER Network Office. 2004. Vision for the 21st Century Information Environment in Ecology (Ecoinformatics) (slide presentation available online: http://slideplayer.com/slide/6950679/).

Appendix B. Data Citation Standards

Data Citation standards for non-USGS data. (Source: USGS Data Management, Data Citation). The following example of a data set citation is from the Earth Science and Information Partners (ESIP).

Zwally, H.J., R. Schutz, C. Bentley, J. Bufton, T. Herring, J. Minster, J. Spinhirne, and R. Thomas. 2003. *GLAS/ICESat L1A Global Altimetry Data V018, 15 October to 18 November 2003*. National Snow and Ice Data Center. data set accessed 2011-07-21 at doi:10.3334/NSIDC/gla01.

A Typical Data Citation Format

- Core required elements of data citation
  - Author or Principal Investigator: The data creator.
  - Release Date/Year of publication: The year of release for a completed data set.
  - Title of data source: The formal title that should generally describe the data set.
  - Version/Edition number: The version of the data set used in the publication.
  - Archive and/or distributer: The organization that manages the data, ideally over a long period of time.
  - Locator/identifier: DOI, ARK, etc. [see Preserve > Persistent Identifiers for more information.]
  - Access date and time: An indication of when the data was accessed; data can be changed or modified over time.
- Other elements that can be included if relevant:
  - Format of the data
  - 3rd party producer
  - Subset of the data used
  - Editor or contributor
  - Publication place
  - Data within a larger work

Best Practices to Support Data Citation

- Assign persistent identifiers with your data sets.
  - If possible, assign a new identifier with each new version of data set.
- Use applications that support metadata creation for your data set.
  - Good metadata associated with a data set is important for access and potential reuse.
  - Examples of metadata applications:
  - Metavist, Morpho, Mermaid, See Describe > Metadata under "Tools" for more information.
- Use standardized keywords that describe your data
  - Allows the data set to be more easily discovered and located
  - Resources for good keywords
    - Biocomplexity Thesaurus (USGS)
    - Global Change Master Directory (NASA)
- Archive the data set with journal publishers and data repositories during the publication process
- When citing a data set in a paper:
  - Use the citation style required by the editor or publisher. If there is no standard, follow a typical format and adapt it to match the style for textual publications.

    - Notify the data repository that holds the data set so they can link to the data set in your paper.
- Encourage other data producers to cite their data sets and make their data available for reuse.

Examples of Data Attribution and Citation

- Bathymetry off the east coast. GIS wetlands layers; everyone does their layer the same way, adds on to the original layer.
- Minerals management services, publish royalties for each state. Links below for minerals management:
  - Assessment of undiscovered continuous oil and gas resources in the Monterey Formation, Los Angeles Basin Province, California, 2015
  - Assessment of undiscovered oil and gas resources of the Mississippian Sunbury shale and Devonian–Mississippian Chattanooga shale in the Appalachian Basin Province, 2016
  - Continuous (unconventional) oil and gas resources, 2000 to 2011, Suggested Citation:

U.S. Geological Survey U.S. Continuous Resources Assessment Team, 2015, U.S. Geological Survey assessments of continuous (unconventional) oil and gas resources, 2000 to 2011: U.S. Geological Survey Digital Data Series DDS-69-MM, 46 p. https://dx.doi.org/10.3133/ds69MM. ISSN: 2327-638X (online)

Ray Obuch: The Alaskan Organic Geochemical Data Base (AOGDM) has an example of a readme file to explain the metadata for the Database: https://pubs.usgs.gov/dds/dds-059/#About

Cassandra Ladino: A good example of an aggregated database that's updated continuously is the Nonindigenous Aquatic Species https://nas.er.usgs.gov/. Here's the map version of database - https://nas.er.usgs.gov/viewer/omap.aspx

Mary Sommer: Department of Interior, Wildland Fire Management (DOI-OWF).

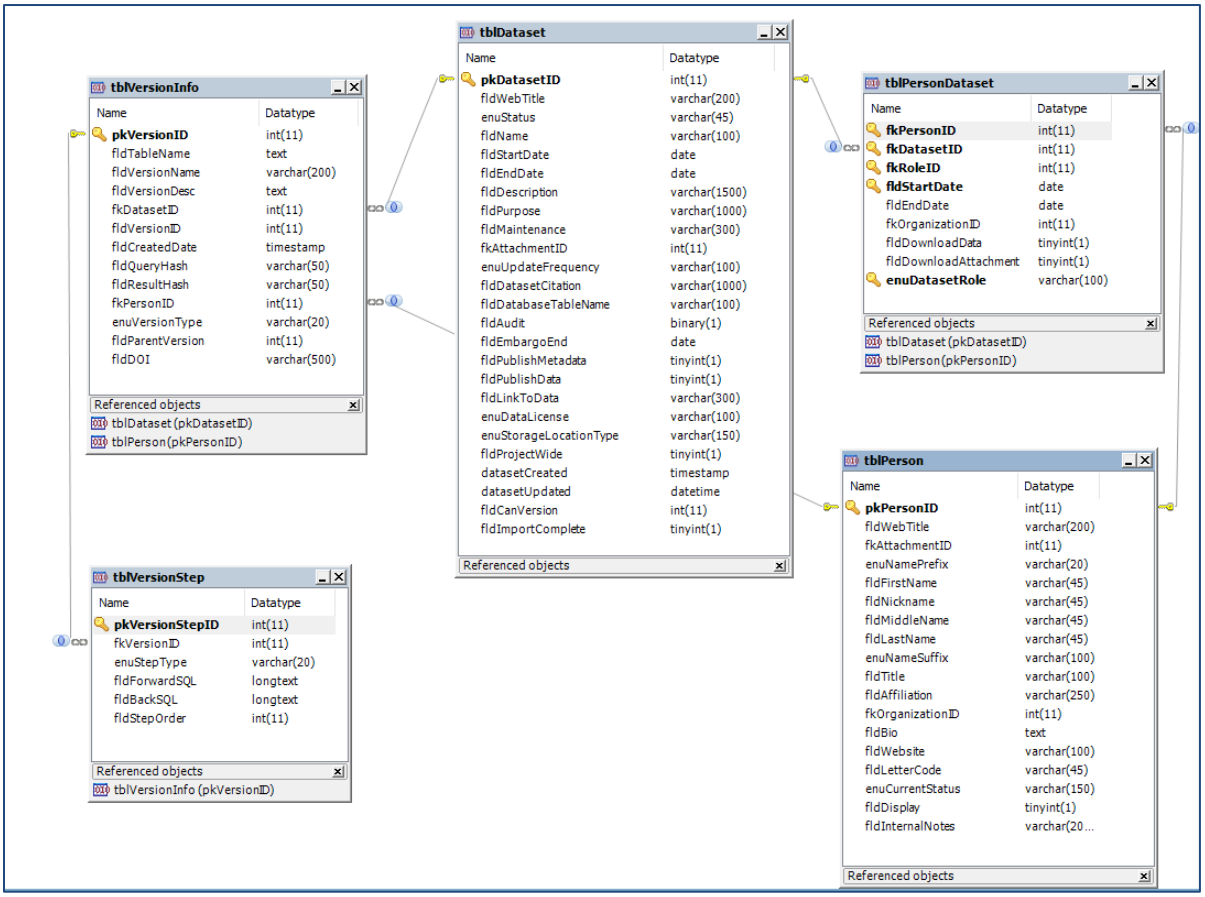
We create data model diagrams then document the data elements within each diagram in a data dictionary. This is a standard approach for most non-government and many government organizations. Or, you can skip the data models and just document your data elements (data attributes) in a spreadsheet. For each entry you would include the following information (you may not need the highlighted ones):

data element name	data properties including: length, data type, case sensitivity, format, example
data element abbreviation	
also known as	
description	
business rules	
data sensitivity	
status	
data steward	
valid values	
source record	
authoritative data source	
system of record	
data stewardship group	
data custodian	

We at DOI-OWF organize this information into formal National Wildfire Coordinating Group (NWCG) Data Standards and post them online. The goal is to have clear definition and understanding of all our data elements. The best source for info about overall Data Management is DAMA (www.dama.org).

Appendix C. Database Schema Example

Examples of database structures that support machine generated citations from data set attributes. These examples are portions of an implementation and are not comprehensive.



Courtesy: James Duncan, [Forest Ecosystem Monitoring Cooperative](#).

A simplified author table follows (courtesy of Raymond Obuch, USGS):

Sample Structure of a data creator/author lookup table

AUTHORID (Unique Primary Key) Format I5		Required	Range of Values 00001-99999
AUTHORID becomes a Foreign Key in the master database header table			
ORGCODE (Foreign Key) Format C10		*Required*	*Range of Values [Az-Zz.0-9]*
ORGCODE (Foreign Key from the Organization Table)			
AUTHORNAME	Format Char30	Required	Not null [Aa-Zz,' ',',]
AUTHORADDRESS	Format Char70	Required	Not null [0-9,Aa-Zz,',',',' ']
AUTHORPHONE	Format Char12	Required	Not null [0-9,-]
AUTHOREMAIL	Format Char25	Required	Not null [Aa-Zz,0-9,@,'.',',_']
000001,GS001RM0005,Raymond Obuch,USGS Box 25046, MS 939, Denver, CO 80225,303-236-5729,obuch@usgs.gov			

If people query the master database, and also select the author, they can pull information from the author id lookup table. You only populate the master table with the authorid number during data entry or database rollup when it occurs.

Appendix D. Glossary

The source for these terms is the CASRAI dictionary developed and maintained by Research Data Canada's (RDC) Standards & Interoperability Committee (http://www.rdc-drc.ca) in collaboration with CASRAI http://dictionary.casrai.org/Category:Research_Data_Domain, unless otherwise noted. The dictionary is made publicly available under a Creative Commons Attribution Only license (CC-BY). The dictionary is in a public comment period through June 30, 2017 and includes many more terms than listed here. Accessed 20170614.

Architecture: Fundamental organization of a system embodied in its components, their relationships to each other and to the environment, and the principles guiding its design and evolution. The term is not always used in normative or prescriptive ways. In some cases, the architecture may need to be flexible and thus more of an open framework rather than being a fixed set of components and services equal to everyone.

At-risk data: Data that are at risk of being lost. At-risk data include data that are not easily accessible, have been dispersed, have been separated from the research output object, are stored on a medium that is obsolete or at risk of deterioration, data that were not recorded in digital form, and digital data that are available but are not useable because they have been detached from supporting data, metadata, and information needed to use and interpret them intelligently.

Authoritative Data: These are officially recognized data that can be certified and provided by an authoritative source.

Authoritative Data Source: Authoritative Data Source (ADS). This is not a source of data, but it is mistakenly used as one when data sources are discussed. ADS is an information technology (IT) term system designers use to identify a system process that ensures the veracity of data sources when a database is created.

Authoritative Source. This is an entity authorized by a legal authority to develop or manage data for a specific business purpose. The data this entity creates are authoritative data.
[source for previous three entries: http://www.iaao.org/uploads/stage.pdf and Ponzio 2004]

Change log: Tracks the progress of each change from submission through review, approval, implementation and closure. The log can be managed manually by using a document or spreadsheet, or it can be managed automatically with a software or Web-based tool.

Citable data: A type of referable data that has undergone quality assessment and can be referred to as citations in publications and as part of research objects.

Comma separated values: A file that contains the values in a table as a series of ASCII text lines organized so that each column value is separated by a comma from the next column's value and each row starts a new line. SYNONYM. **CSV**. RELATED TERM. Character separated values

Consensus standard: Developed through the cooperation of all parties who have an interest in participating in the development and/or use of the standards. Consensus requires that all views and

objections be considered, including feasibility, and that an effort be made toward their resolution. Consensus implies more than the concept of a simple majority, but not necessarily unanimity. Consensus standards should be viewed as minimum acceptable standards, not an ideal or maximum target objective.

Data attribution: 1) Acknowledge or credit the person or organization that created, validated or curated a data set. 2) Attach characteristics (metadata attributes) to data sets that are machine readable, linkable to the data set as an object, and ensure persistence of the data set as it was first used.

Data citation: Offers proper recognition to authors as well as permanent identification through the use of global persistent identifiers in place of URLs which can change frequently. Use of universal numerical fingerprints (UNFs) guarantees to the scholarly community that future researchers will be able to verify that data retrieved is identical to that used in publication decades earlier, even if it has changed storage media, operating systems, hardware, and statistical program format. Data citation is provided in a similar way that researchers routinely include bibliographic references to traditionally published resources. Data citation should include the following elements: (a) Name Principal Investigator/Author/Data Creator; (c) Release Date/Year of Publication - year of release, for a completed data set; (d) Title of Data Source - formal title of the data set; (e) Version/Edition Number - the version of the data set used in the study; (f) Format of the Data - physical format of the data; (g) 3rd Party Data Producer - refers to data accessed from a third party repository; (h) Archive and/or Distributor - the location that holds the data set; (i) Locator or Identifier - includes Digital Object Identifiers (DOI), Handles, Archival Resource Key (ARK), etc.; (j) Access Date and Time - when data are accessed online; (k) Subset of Data Used - description based on organization of the larger data set; (l) Editor or Contributor - reference to a person who compiled data, or performed value-added functions; (m) Publication Place - city, state, and country of the distributor of the data; and, (n) Data within a Larger Work - refers to the use of data in a compilation or a data supplement (such as published in a peer-reviewed paper).

Data dictionary: A collection of descriptions of the data objects or items in a data model. A first step in analyzing a system of objects with which users interact is to identify each object and its relationship to other objects. This process is called data modeling and results in a picture of object relationships. After each data object or item is given a descriptive name, its relationship is described (or it becomes part of some structure that implicitly describes relationship), the type of data (such as text or image or binary value) is described, possible predefined values are listed, and a brief textual description is provided. This collection can be organized for reference into an eBook called a data dictionary.

Data governance: The exercise of authority, control and shared decision making (planning, monitoring and enforcement) over the management of data assets. It refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures, and a plan to execute those procedures.

Data identifier: An identifier that uniquely distinguishes one set of data from all others. Examples include: Archival Resource Key (ARK); Digital Object Identifiers (DOI); Extensible Resource Identifier (XRI); HANDLE; Life Science ID (LSID); Object Identifiers (OID); Persistent Uniform Resource Locators (PURL); URI/URN/URL; (Universally Unique Identifier) UUID.

Data management plan: A formal statement describing how research data will be managed and documented throughout a research project and the terms regarding the subsequent deposit of the data with a data repository for long-term management and preservation.

Database: A collection of data that is organised in a according to a conceptual structure/model describing the characteristics of these data and the relationships among their corresponding entities, supporting one or more application areas. A database allows its contents to be easily accessed, managed and updated. The type of database used depends on the requirements of the study. A common type is the relational database, where data are related to each other in a systematic manner so that they can be reorganised and accessed in a number of different ways. A database may house one or many datasets.

Dataset: Any organized collection of data in a computational format, defined by a theme or category that reflects what is being measured/observed/monitored. The presentation of the data in the application is enabled through metadata.

FAIR: FAIR Principles address the lack of widely shared, clearly articulated, and broadly applicable best practices around the publication of scientific data. FAIR Principles address these needs by providing a precise and measurable set of qualities a good data publication should exhibit - qualities that ensure that the data are **Findable, Accessible, Interoperable, and** Reusable (FAIR) (Wilkinson *et al.* 2016).

Information silos: Heterogeneous data sources.

Interoperability: The capability to communicate, execute programs, or transfer data among various functional units in a useful and meaningful manner that requires the user to have little or no knowledge of the unique characteristics of those units. Foundational, syntactic, and semantic interoperability are the three necessary aspects of interoperability.

ISO 19115 Metadata profile: A metadata profile that specifies the elements and syntax to be used when implementing the international geospatial standard (ISO 19115: 2003) in North America. SYNONYM. North American Profile for ISO 19115; NAP.

Manage data sets in a repository: Implement the policies that govern the arrangement, naming, descriptive metadata, provenance metadata, representation metadata, administrative metadata, access controls, retention, disposition, integrity, and replication of digital objects.

Minimal metadata: A description with very little curation that would include at least a name and PID of a data object. Minimal metadata are only marginally targeted at discovery since there is much better infrastructure to accomplish this.

Persistent identifier: A persistent identifier is a long-lasting reference to a digital object that gives information about that object regardless what happens to it. Developed to address "link rot," a persistent identifier can be resolved to provide an appropriate representation of an object whether that objects changes its online location or goes offline. SYNONYM. PID

Provenance metadata: Information concerning the creation, attribution, or version history of managed data. Provenance metadata that indicates the relationship between two versions of data objects and is generated whenever a new version of a data set is created. Examples include: (i) the name of the program that generated the new version, (ii) the commit id of the program in a code version control system like GitHub, (iii) the identifiers of any other data sets or data objects that may have been used in creating the new version. Provenance information is gathered along the data lifecycle as part of curation processes. A finer level of provenance metadata would be concerned only with data flowing between various stores such as curated databases and managed repositories. Provenance metadata are designed to allow queries over the relationship between versions and includes either or both fine-grained and coarse-grained provenance data. Different applications may store different provenance data.

Reproducible research: Published results can be replicated using the documented data, code, and methods employed by the author or provider without the need for any additional information or needing to communicate with the author or provider. SYNONYM. Reproducibility

Repurposed data: New data sets obtained by combining data appropriately from a variety of existing files, generating new data products that did not previously exist. Repurposed data result from data wrangling. RELATED TERM. Data wrangling

Relational database: A collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. The standard user and application program interface to a relational database is the structured query language (SQL). SQL statements are used both for interactive queries for information from a relational database and for gathering data for reports. In addition to being relatively easy to create and access, a relational database has the important advantage of being easy to extend. After the original database creation, a new data category can be added without requiring that all existing applications be modified. A relational database is a set of tables containing data fitted into predefined categories. Each table (which is sometimes called a relation) contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns. For example, a typical business order entry database would include a table that described a customer with columns for name, address, phone number, and so forth. Another table would describe an order: product, customer, date, sales price, and so forth. A user of the database could obtain a view of the database that fitted the user's needs. For example, a branch office manager might like a view or report on all customers that had bought products after a certain date. A financial services manager in the same company could, from the same tables, obtain a report on accounts that needed to be paid. When creating a relational database, the domain of possible values in a data column can be defined as well as

further constraints that may apply to that data value. For example, a domain of possible customers could allow up to ten possible customer names but be constrained in one table to allowing only three of these customer names to be specifiable.

Schema: 1. The organization or structure for a database. The activity of data modeling leads to a schema. (The plural form is schemata.) The term is used in discussing both relational databases and object-oriented databases. The term sometimes seems to refer to a visualization of a structure and sometimes to a formal text-oriented description. Two common types of database schemata are the star schema and the snowflake schema. 2. A formal expression of an inference rule for artificial intelligence (AI) computing. The expression is a generalized axiom in which specific values or cases are substituted for each symbol in the axiom to derive a specific inference.

Structured data: Data whose elements have been organized into a consistent format and data structure within a defined data model such that the elements can be easily addressed, organized and accessed in various combinations to make better use of the information, such as in a relational database. SYNONYM. Structured information

Unstructured data: Data that have not been organized into a format and identifiable data structure that makes them easy to access and process. These data can often be searched as long as they are digital, but they are difficult to use for computer analyses. SYNONYM. Unstructured information

Use case: A methodology used in system analysis to identify, clarify, and organize system requirements. The use case is made up of a set of possible sequences of interactions between systems and users in a particular environment and related to a particular goal. It consists of a group of elements (e.g. classes and interfaces) that can be used together in a way that will have an effect larger than the sum of the separate elements combined. The use case should contain all system activities that have significance to the users. A use case can be thought of as a collection of possible scenarios related to a particular goal, indeed, the use case and goal are sometimes considered to be synonymous.

Version control: Control over time of data, computer code, software, and documents that allows for the ability to revert to a previous revision, which is critical for data traceability, tracking edits, and correcting mistakes. Version control generates a (changed) copy of a data object that is uniquely labeled with a version number. The intent is to track changes to a data object, by making versioned copies. Note that a version is different from a backup copy, which is typically a copy made at a specific point in time, or a replica. SYNONYM. Source control; Revision control; Versioning. RELATED TERM. Universal numeric fingerprint; Data citation