esa ECOSPHERE

SPECIAL FEATURE: EMERGING TECHNOLOGIES IN ECOLOGY

Data system design alters meaning in ecological data: salmon habitat restoration across the U.S. Pacific Northwest

Stephen L. Katz,[1],† Katie A. Barnas,[2] Monica Diaz,[2] and Stephanie E. Hampton[3]

[1]*School of the Environment, Washington State University, Pullman, Washington 99164 USA*
[2]*Northwest Fisheries Science Center, NOAA Fisheries Service, Seattle, Washington 98112 USA*
[3]*Center for Environmental Research, Education and Outreach, Washington State University, Pullman, Washington 99164 USA*

Citation: Katz, S. L., K. A. Barnas, M. Diaz, and S. E. Hampton. 2019. Data system design alters meaning in ecological data: salmon habitat restoration across the U.S. Pacific Northwest. Ecosphere 10(11):e02920. 10.1002/ecs2.2920

Abstract. As an increasing variety and complexity of environmental issues confront scientists and natural resource managers, assembling the most relevant and informative data into accessible data systems becomes critical to timely problem solving. Data interoperability is the key criterion for succeeding in that assembly, and much informatics research is focused on data federation, or synthesis to produce interoperable data. However, when candidate data come from numerous, diverse, and high-value legacy data sources, the issue of data variety or heterogeneity can be a significant impediment to interoperability. Research in informatics, computer science and philosophy has frequently focused on resolving data heterogeneity with automation, but subject matter expertise still plays a large role. In particular, human expertise is a large component in the development of tools such as data dictionaries, crosswalks, and ontologies. Such representations may not always match from one data system to another, presenting potentially inconsistent results even with the same data. Here, we use a long-term data set on management actions designed to improve stream habitat for endangered salmon in the Pacific Northwest, to illustrate how different representations can change the underlying information content in the data system. We pass the same data set comprised of 49,619 records through three ontologies, each developed to address a rational management need, and show that the inferences drawn from the data can change with choice of data representation or ontology. One striking example shows that the use of one ontology would suggest water quality improvement projects are the rarest and most expensive restoration actions undertaken, while another will suggest these actions to be the most common and least expensive type of management actions. The discrepancy relates to the origins of the data dictionaries themselves, with one designed to catalog management actions and the other focused on ecological processes. Thus, we argue that in data federation efforts humans are “in the loop” rationally, in the form of the ontologies they have chosen, and diminishing the human component in favor of automation carries risks. Consequently, data federation exercises should be accompanied by validations in order to evaluate and manage those risks.

Key words: applied epistemology; bioinformatics; crosswalk; data confederation; data federation; data synthesis; ecoinformatics; interoperability; ontology; restoration; Special Feature: Emerging Technologies in Ecology.

Received 23 July 2019; accepted 6 September 2019. Corresponding Editor: Judy Cushing.

† **E-mail:** Steve.Katz@WSU.edu

Introduction

Addressing real-world natural resource management challenges necessitates matching the most salient information to the needs of decision makers. Unfortunately, the most salient data are often dispersed among distributed data holders, making it difficult to get the best data in front of the right people (Katz et al. 2007, Volk et al. 2014, Williams and Labou 2017). This broad distribution of data may not be a problem, but rather may develop rationally, and for multiple reasons. For example, many managed natural resources span broad geographic areas that fall under the domains of numerous agencies, each with different needs and practices (Bayer 2006). Further, new science and management questions often emerge on a larger geographic or conceptual scale than when individual natural resource monitoring programs were funded, designed and executed (Carpenter et al. 2009). In many of these cases, data were a means for an individual or organization to answer a particular question, rather than a research product to be protected, curated, or distributed. After collection, these data become part of the dark data, largely unknown and inaccessible outside of the organization where they originated (Heidorn 2008). In response, many groups and individuals increasingly see the need to combine disparate data into a synthetic, relevant data corpus (Romanello et al. 2005, Jones et al. 2006, Reichman et al. 2011).

Not only are these data often highly distributed, but consequent to their independent development, they often lack standardized methods and formats (Heidorn 2008, Brunt and Michener 2009, Vassiliadis and Simitsis 2009, Kolb et al. 2013, Theodorou et al. 2014), creating a substantial challenge in simply discovering, let alone combining, the vast pool of data that result from largely public investments in natural resources science. Creating a single, functional data set from disparate source data has been termed data (con)federation, synthesis, or integration (Hull 1997, Haas et al. 2002, Romanello et al. 2005, Reichman et al. 2011), and in the domain of corporate or commercial data, some data federation workflows have been characterized as extraction-transformation-loading, or ETL (Shu et al. 1977, Vassiliadis and Simitsis 2009, Theodorou et al. 2014). In an ETL workflow, data extracted from source data are cleaned or otherwise transformed into a common format, screened for compliance with a uniform data standard, and loaded into a single data warehouse (Vassiliadis and Simitsis 2009).

Data heterogeneity forms a significant challenge to successful data federation and can exist at several levels of organization within data systems (Qian 1993, Hammer and McLeod 1999, Vassiliadis and Simitsis 2009). Data heterogeneity is often confronted as a characteristic of the data representation, such as differences in units or confusion arising from common terms meaning different things (=homonyms) or multiple terms referring to the same thing (=synonyms; Vassiliadis and Simitsis 2009). Consequently, data heterogeneity is often addressed as a data quality issue where heterogeneity is an undesirable property (Dong and Naumann 2009). However, differences among metadata may or may not overlay real semantic or ontological differences among the entities collecting, curating, and contributing data to a synthetic corpus, and this is often particularly challenging when data are represented as strings, text, and narrative (Vassiliadis and Simitsis 2009). While there are numerous approaches to addressing this issue, it is still not clear how the methods for resolving differences among multiple, (sometimes highly) heterogeneous metadata catalogs may preserve or alter the relationships among data elements, the underlying informational content in the data, and ultimately the interpretation of the data during analysis.

The key technical development to empower large-scale data federation projects is tools to resolve semantic differences among disparate source data. These include data dictionaries, crosswalks, and ontologies. Historically, data federation in ecology was conducted by individual researchers often exercising judgment based on domain expertise, limiting the scope and rate of data federation. It is anticipated now that the greater scientific community is on the cusp of a transition where automation, leveraging machine learning and algorithmic approaches, will increase the velocity and accessibility of interoperable, federated data (Qian 1993, Hammer and McLeod 1999, Haas et al. 2002).

These emerging tools for semantic integration come with characteristics and features that can impact their behavior and utility. So as we deploy crosswalks and ontologies to federate ecological data, we need to be circumspect in their use just as we would with other emerging technologies. Here, we present an examination of this issue using a case study to demonstrate how the information content within a single, large natural resource management data set responds when crosswalked using different ontologies. We start with a short clarification of what we mean by data dictionaries, crosswalks, and ontologies. We then present an illustrative case study of a data system curating over 40,000 records collected over the last 20 yr of habitat restoration projects in the Pacific Northwest. The case study amounts to passing the same base data through three distinct ontologies to show how each alters the information content of the data. We conclude with a discussion of the impact of this behavior, and the implications for changes in information content resulting from the federation process. Importantly, identifying these behaviors is not intended to be a critique of these tools, as these behaviors are properties that are anticipated based on semantics and representation theory, not dependent on data quality per se, and are therefore generic (Hull 1997, Kuhn 2003, Obrst 2003). Rather, we think it a useful discussion to explore the behavior of these rapidly developing tools, and their philosophical underpinnings, so that we can better evaluate their robustness and ultimately maximize their utility.

Data Dictionaries, Ontologies, and Crosswalk Translators

Data dictionaries constitute a form of metadata that describes the informational content of fields within a data set, and so defines the fields in a database to allow classification of the data into groups with like properties. High performing dictionaries provide users with a clear discrimination of similar observations within a data set, and sorting of observations that are unlike. More practically, data dictionaries classify data into a finite number of well-defined fields despite variability in the original source data. Variability can take many forms. For example, spatial location for a management action can exist in numerous formats such as latitude/longitude and township-range-section. Standardized definitions allow for rapid manipulation, comparison, and communication of data content and are thus a key prerequisite for data interoperability.

When we apply analysis to data sets, those data sets are a representation of the world that the data collection was intended to describe. In the case of habitat restoration projects, each action on the ground is represented in a data set by a set of variables that include project type, location, duration, and extent of habitat restored, which is clearly capturing many but not all facets of that action. Thus, such representations are inherently less complex than real life and are thus abstractions of it. If functioning well, the data set design captures enough of the real world to address the original motivating questions. With restoration projects for example, an effective data system would need to capture enough of the restoration process to test hypotheses about habitat impacts, ecosystem responses, distribution of restoration resources, or other relevant questions. If they are to be effective, representations of the real world are also expressed as ontologies, which capture not just the definitions of the fields, in the form of a data dictionary perhaps, but also their conceptual connections or internal logic in a sufficiently explicit manner for the information to be sharable among users (Gruber 1993, Uschold and Gruninger 2004). The conceptual connections can take the form of explicit relationships between data elements, such as is-a or has-dimensions-of (Madin et al. 2007) but can also be manifest simply in the topology of connections or hierarchies within a data system (Gruber 1993). In practice, data system development has sometimes proceeded without being conscious of the ontology, but once the language and topology are specified, the ontology will manifest, along with the consequences of diversity among different ontologies.

Like any representation, the process of database design or ontology development can be biased by the perspective, needs, and background of the designer of the data system (Gruber 1993). When multiple, independent but similar efforts develop data systems, differences among these efforts can produce a different mapping of the world based on distinct data dictionaries. If our ambition is subsequently to combine data from different sources, we must at least resolve differences among source data dictionaries, including potential underlying conceptual differences, with a translation or crosswalk.

A crosswalk is described as a set of transformations that map the content of a source metadata standard onto analogous elements of a target metadata standard (Pierre and LaPlant

2000). The implications of crosswalk data mapping include the need to map both semantic and structural aspects of the data sets and the challenge of topological differences. Topological differences could include one data set being hierarchical, but the other not (Pierre and LaPlant 2000), which is an important issue for crosswalks, and in fact manifests in our test case.

Crosswalks function as translators, and so are subject to the limitations common to other translation mechanisms. There is certain to be loss of functional information in the process of translating that information into a representation, or from one representation to another. This underdetermination of information by data has been referred to as a class of problems in philosophy called the indeterminacy of translation (Quine 1970). Although somewhat contentious in the philosophy literature (Schick 1973, Wright 2017), it is a relatively common experience in informatics and statistics.

Case Study: Habitat Restoration Data from the Pacific Northwest Passed Through Multiple Ontologies

In the Pacific Northwest, the listing of five Pacific Salmon species as threatened and endangered across much of Washington, Oregon, and Idaho motivated large investments in habitat management that have at times neared 400 Million U.S. Federal dollars per annum in the Columbia River Basin alone (G.A.O. 2002). In particular, improvements in fish habitat via restoration, and the anticipated improvement in fish survival, have been widely applied as a principal management tactic to compensate for the increased mortalities that led to salmon declines and ESA listing (Kareiva et al. 2000). Initially, regional investments in restoration actions were not documented with a targeted or coordinated implementation monitoring program, making evaluations of project effectiveness impossible (G.A.O. 2002; Katz et al. 2007). There were, however, numerous and disparate data sets of restoration actions curated by project sponsors and funding agents. Over time, these diverse agents have documented their projects for a variety of reasons including justification of expenditures by funders, management or regulatory requirements, and research needs. Each agency or organization covered a subset of the total spatial scale of endangered salmon management, and applying all the data to regional salmon management required federation of all the data sets.

In response, our group embarked on an ongoing synthesis of the available project-level data on restoration in the Pacific Northwest (Katz et al. 2007). We aimed to census all habitat restoration projects that could impact salmon habitat in the states of Washington, Oregon, Idaho, and Montana. At its initiation, we hoped that the resulting data system could be used to evaluate the impact of habitat restoration on improvements in endangered fish survival. The resulting Pacific Northwest Salmon Habitat Project (PNSHP) database is described in detail elsewhere (Katz et al. 2007, Barnas and Katz 2010, Hamm 2012, Barnas et al. 2015), and while that project informed restoration practice, it also revealed some important insights into the broader challenges of data federation.

In assembling the PNSHP database, a census of existing habitat restoration project data consisted of the following steps:

1. Obtain all available data from as many sources as possible in formats that ranged from relational electronic databases to paper project files;
2. Manually read the data to determine the attributes of each data set of interest and how they vary from source to source;
3. Resolve the common information encoded in each of the data systems based on associated source metadata and consultation with subject area experts, and then resolve the diversity of project data into a common descriptive metadata standard;
4. Develop crosswalks between each source data attribute definitions and a common data dictionary for the final resolved data system.

This approach resulted in the importation of 26 different data sets with varying formats held by private, local, state, tribal, and federal entities into one format and machine-readable database, the PNSHP database. Two of the 26 data sets were compiled by individuals contacting local and municipal land agents to obtain 72 records for individual projects, similar to a door-to-door canvas. These agents collected data from diverse individuals, but with a common data dictionary to record project records resulting in a total of 96 unique sources, but among them there were only 25 independent data systems to be reconciled.

Based on the available data, the final PNSHP data structure of the output data system includes fields for project identifier, sponsor, contact info, location, project type, timing, cost, goals, and existence of monitoring. As described in Katz et al. (2007), at the time there was no single standard format for any of these attributes requiring some degree of reconciliation for all of them. However, we found the greatest degree of heterogeneity in the description of project type, and for this case study, we will focus on project type to evaluate the effect of different data dictionaries on the mapping of information onto a data structure in our analysis.

The PNSHP database defined the restoration actions in a hierarchical scheme. In all levels in this hierarchy, projects are defined by the actions taken by the project sponsor rather than other possible bases for definition; this basis establishes the internal logic of this ontology. First, project subtype was defined based on the action taken at the worksite (e.g., culvert installation, culvert replacement, fish by-pass installation). Second, project subtypes were aggregated into project types (e.g., barrier removals). This hierarchical scheme emerged from steps 2 and 3 in the process outline above; the raw data from the 26 possible data holders were inconsistent in describing both type and subtype of project, but given that a project type could be deduced from its subtype, this scheme allowed the bottom-up capture of the greatest number of project records. In addition, it also captured the greatest diversity of language used by the project sponsors themselves. From this process, we were also able to generate a single data dictionary for project type as described in Katz et al. (2007) and in Appendix S1: Table S1. Resolving the project records from each data contributor into that single data dictionary also produced a series of crosswalks for each of the 26 unique source data designs that translated the attribute descriptors to the output data set.

At the same time that we were assembling PNSHP data (mid-2000s), two other, parallel projects were underway with similar goals, but different organic questions, and internal logics. PNSHP collaborated with both of these other restoration compilations, necessitating semantic differences among the three projects be resolved, and crosswalks created. The crosswalks we created between the data dictionaries of these three databases allow us now to evaluate the impact of differences among the ontologies on the mapping of information onto data.

The National River Restoration Science Synthesis (NRRSS) was a data synthesis project through the National Center for Ecological Analysis and Synthesis, which sought to evaluate the state of practice of stream restoration by compiling restoration data for eight regions (or nodes) across the Unites States. The motivation was to identify successful demonstrations of different types of stream restoration and in so doing highlight the reasons for their success. National River Restoration Science Synthesis grouped similar projects into intents based on the ecological process goals of the restoration action distinct from the action itself, for example, floodplain reconnection (Bernhardt et al. 2005). National River Restoration Science Synthesis did not include a sub-level in their intents and thus was not hierarchical. The NRRSS intents included four categories that existed in other parts of the country but were either not represented in the Pacific Northwest or they did not impact salmon habitat. These included instream species management, out of stream land acquisitions, non-water quality related storm water projects, and narrowly defined aesthetic, recreational, or educational actions. Consequently, the NRRSS data system had 13 intents, but only nine were represented in this crosswalk exercise. PNSHP data were integrated into the NRRSS national-scale database and at that time constituted 77.9% of the national total of project records (Bernhardt et al. 2005).

At the same time, the Pacific Coast Salmon Recovery Fund (PCSRF) began capturing data on PCSRF-funded management actions related to salmon in the western United States. The PCSRF was established by congress in 2000 as a means by which NOAA Fisheries Service would provide grants to States and Tribes to assist State, Tribal, and local salmon conservation and recovery efforts. Originally, the formation of PCSRF was requested by the governors of the States of Washington, Oregon, California, and Alaska in response to Endangered Species Act (ESA) listings of West Coast salmon and steelhead populations. The desire to assemble a comprehensive database of project records came later in order to support annual reports to the U.S. Congress to justify spending of PCSRF funds. The context of accountability in the PCSRF process resulted in projects being categorized in part by where they occur on the landscape (Estuarine, Wetland, Upland, Instream, etc.) with jurisdictions and constituent interest in mind. Some additional project types, such as fish exclusion screens, were characterized by the action taken rather than the location. However, actions like fish screens are also distinct in being tied to specific land uses where water withdrawals are occurring (Barnas et al. 2015), and so there is at least some implicit location information that structured the project type designations in PCSRF.

In all three cases, the project included the defined vocabulary in the context of an underlying conceptualization of the project goals—for example, what the action is, what it is intended to achieve ecologically, where it is performed, respectively (summarized in Table 1). As such, each of these representations needs to be seen not just as a dictionary of terms or vocabulary, but also as an ontology (Gruber 1993, Chandrasekaran et al. 1999).

To evaluate the impact of different ontologies on the information content of data systems, we compare the patterns in data that result when the same raw data (i.e., project records as reported in Katz et al. 2007) are represented in three different data systems that rely on their own ontology. For this case study, we focus on two patterns in the data. We first ask how the distribution of project types changes when different ontologies are used, and in particular whether our original hypothesis about cost determining the distribution of restoration actions is altered by the choice of ontology. Second, we ask whether the correlation between the distribution of restoration projects and ecological need varies based on the choice of ontology. The source data for project records and cost are from the PNSHP database that is publically available (https://www.webapps.nwfsc.noaa.gov/pnshp).

Ecological need is expressed as ecological concerns that are defined in Hamm (2012), where a mapping was developed between restoration action and the ecological concern it can address, currently in use by NOAA fisheries for habitat assessments for salmon recovery. The data on

Table 1. Summary of characteristics of the contributing data systems to the ontology comparison.

Data system	PNSHP	NRRSS	PCSRF
Name	Pacific Northwest Salmon Habitat Project data base	National River Restoration Science Synthesis	Pacific Coast Salmon Recovery Fund
Contact	https://www.nwfsc.noaa.gov/research/divisions/cb/mathbio/salmon_habitat.cfm	https://www.nceas.ucsb.edu/riverrestoration/	https://www.webapps.nwfsc.noaa.gov/pcsrf
Funding Source	U.S. Federal Government: National Oceanographic and Atmospheric Administration	Academic/NGO partnership: National Center for Ecological Analysis and Synthesis and American Rivers	U.S. Federal Government: National Oceanographic and Atmospheric Administration
Context for developing data system	Supporting effectiveness monitoring for restoration	Assess patterns of restoration practice and explore indicators of restoration success	Assessment of accountability for funds applied to restoration actions.
Basis for classifying project types (i.e., internal logic)	Action taken at the worksite	Ecological process affected by restoration	Location of action on landscape
Hierarchical typology of restoration types	Yes	No	No
Number of project types present	11	9	9
Total captured	45,073	32,025 (plus "ULUM" & "WL"=45,073)	35,056 (plus ULUM & F = 45,073)

Notes: ULUM, upland land use management. Total number of records is 49,619.

ecological concerns were amassed in an inventory of ESA recovery plans from across Washington, Oregon, and Idaho and are documented in Barnas et al. (2015). In reality, this final step amounts to three separate ontologies of restoration actions being correlated with ecological concern these actions are believed to address. This applied example illustrates how crosswalked data can inform management decisions about what type of restoration to pursue, and also how the crosswalk itself can influence the outcome. For all questions, correlations among distributions are compared with Kendall's tau (τ) rank correlations (Sokal and Rohlf 1969) and calculated in the R environment for statistical computing (Ver. 3.4.4, R. Core Team 2013).

Results

The total number of restoration project records in the federated data set was 49,619. We were able to translate most, but not all project records across the three data systems. As reported previously (Katz et al. 2007), 4546 project records did not provide enough information to identify project type (labeled other in the PNSHP data system) and failed to be mapped onto a project type in any of these data systems. In addition, due to differences in the types of restoration captured and excluded by each of the databases, some project records were not crosswalked into all of the data systems compared here. The net project records successfully crosswalked to each of the three data systems are summarized in Table 1. As mentioned above, the PNSHP database uses eleven types, with the breakdown into subtypes creating 86 type–subtype combinations (detailed in Appendix S1: Table S1), while NRRSS and PCSRF each categorize projects as nine specific intents or types (Appendix S1: Tables S2, S3). In addition, the NRRSS data system did not have a category for upland land use management (ULUM in Table 1) and some wetland-based (WL) restoration projects. These amounted to 13,047 project records not captured by the NRRSS system. Similarly, the PCSRF data system did not recognize forestry management and some upland land use management (F and ULUM) types that were not clearly related to salmon recovery funding, totaling 10,025. In the case of the NRRSS system, these projects were not deemed river restoration per se; in the case of PCSRF, large land use projects (distinct from small projects, such as fencing or livestock management) were not commonly funded under the program and tracking them was likewise not a priority. Aside from the indeterminate project records, designated other above, all records were definable, and the difference in number of records between the data systems represents intentioned decisions about what is included and excluded in each data system, rather than ability to define the project records on the basis of the available information.

Numerous project records had multiple project types in PNSHP (6643, 23%) and were resolved into more than one project type category. In these cases, a single project record had data for multiple worksite locations and often different actions at each location (e.g., when a project was both Restore Riparian Function and Sediment Reduction types, it would be counted as a 1 in both type categories). Since cost was only available at the project level, in the cases of a one-to-many relationship between project record and project subtype and/or location existed, the full project cost was attributed to all project type–subtype combinations in that record.

Passing the same project data to each of these three different data dictionaries resulted in three distinct distributions of projects based on type (Fig. 1). Among the commonalities across systems, riparian restoration projects (e.g., Restore Riparian Function, Riparian Habitat Projects, Riparian Management) were the most common project type under all three sets of definitions and yielded similar counts of 5989–6106 and median costs of $9,600–$9,921 (US$ are reported and are undeflated). Upland Management projects were among the three most common types under both the PNSHP and the PCSRF data definitions; however, the number of projects and costs varied by over 11% (PNSHP count 5914, median cost $6,092, PCSRF count 6583, median cost $7,810). Although the net counts were higher when crosswalked to PCSRF definitions, this difference reflects the larger number of project records being reclassified both into and out of the Upland Management category. Numerous sediment reduction projects in PNSHP were classified as Upland Management in PCSRF, but this did not entirely balance a much larger number of

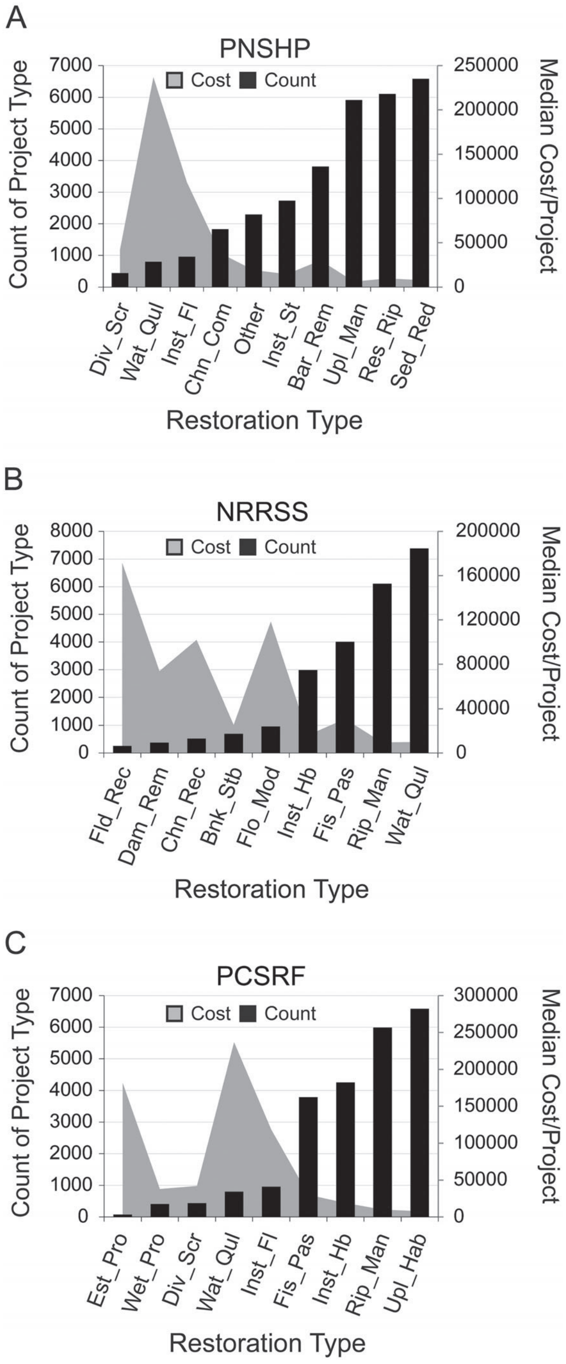

Fig. 1. Histograms of number of restoration projects (bars) and median cost (profile plot) for each category of project. In each plot, the raw data and the dependent variables are the same, but the different data records are assigned to a project category based on the ontology in use by that data synthesis project. Therefore, the labels for the categories (i.e., the *X*-axis) differ somewhat in language; the exact mapping is described in the supplement (Appendix S1: Tables S2, S3). (A) Distribution of projects and costs using the Pacific Northwest Salmon Habitat Project Tracking Database ontology. Abbreviations of restoration types are Div_Scr, Fish Diversion Screens; Wat_Qul, Water Quality Improvement; Inst_Fl, Instream Flow; Chn_Com, Restore Instream Complexity via

(Fig. 1. *continued*)

Channel Complexity; Inst_St, Restore Instream Complexity via Instream Structure; Bar_Rem, Barrier Removal; Upl_Man, Upland Management; Res_Rip, Restore Riparian Function; Sed_Red, Sediment Reduction; and Other is undefinable (see text). (B) Distribution of projects and costs using the National River Restoration Science Synthesis. Abbreviations of restoration types are as follows: Fld_Rec, Floodplain Reconnection; Dam_Rem, Dam Removal; Chn_Rec, Channel Reconnection; Bnk_Stb, Bank Stabilization; Flo_Mod, Flow Modification; Inst_Hb, Instream Habitat Improvement; Fis_Pas, Fish Passage; Rip_Man, Riparian Management; and Wat_Qul, Water Quality Management. (C) Distribution of projects and costs using the Pacific Coastal Salmon Recovery Fund ontology. Abbreviations of restoration types are as follows: Est_Pro, Estuary Projects; Wet_Pro, Wetlands Projects; Div_Scr, Fish Diversion Screens; Wat_Qul, Water Quality Projects; Inst_Fl, Instream Flow Projects; Inst_Hb, Instream Flow Projects; Fis_Pas, Fish Passage Improvement Projects; Rip_Man, Riparian Management Projects; and Upl_Hab, Upland Habitat Projects.

Upland Management projects that were reclassified to a more specific subtype in PNSHP.

Overall, we found distinct variability in the rank of projects by type across the data systems. In one of the biggest contrasts, numerous sediment reduction and water quality improvement projects in the PNSHP were captured under the Water Quality Management intent under the NRRSS definitions. Thus, NRRSS Water Quality Management projects totaled 7382 records, making it the most common intent, with a median cost of $10,000. In PCSRF and PNSHP, Water Quality Projects and Water Quality Improvement Projects totaled only 799 projects but were the most expensive by far at $236,709.

In our original analysis of PNHP data, we found a strong negative correlation between project count and project cost (Katz et al. 2007). We find this pattern repeated when PNSHP is crosswalked into the NRRSS and PCSRF ontologies. Regardless of which ontology, less expensive projects were more commonly performed, although the strengths of the relationships differ. Kendall's rank order correlations between project abundance and cost were stronger for the PNSHP data ($\tau = -0.73$, $P = 0.002$, with nutrient additions excluded, see Katz et al. 2007; and $\tau = -0.69$, $P = 0.005$, PCSRF; and $\tau = -0.6$, $P = 0.03$, NRRSS).

To ask how well the distribution of projects matched the distribution of environmental needs, we performed rank order correlations of frequency of expressed ecological concerns and frequency of projects that address that concern. This metric of match between need and action is crude and contains a heavy bias toward correctly matching action with need as any relevant project is scored as addressing a need regardless of how small the quantitative investments in that project type may be (Barnas et al. 2015). Thus, the point here is not that restoration actions are or are not executed strategically, but rather how the ontology alters the information content in the data.

We find a wide variation in the match between ecological concern and distribution of restoration, and a wide variation in how that match is reflected when the data are expressed in the three different data systems. A histogram of ecological concerns across the extent of salmon recovery reveals that concerns with habitat connectivity and complexity (peripheral and transitional habitats and channel structure and form) are the most commonly expressed ecological concerns but these needs are less commonly addressed with restoration (Fig. 2). The most common restoration actions are designed to control water quality and sediment issues, even though these needs are expressed with intermediate frequency. Similarly, nutrient limitation projects have been prioritized more so than their ecological concern. Conversely, peripheral and transitional habitats, channel structure and form, and habitat quality receive restoration at a lower rate than the frequency with which they are identified as a concern.

The numerical correlations between concern and actions do differ among data dictionaries. Kendall's rank order correlations between relative frequency of projects type and frequency that the concern was stated in recovery plans were $\tau = 0.44$ ($P = 0.12$) for the PNSHP data, and $\tau = 0.08$ ($P = 0.82$) and $\tau = 0.37$ ($P = 0.18$) for the PCSRF and NRRSS data dictionaries,

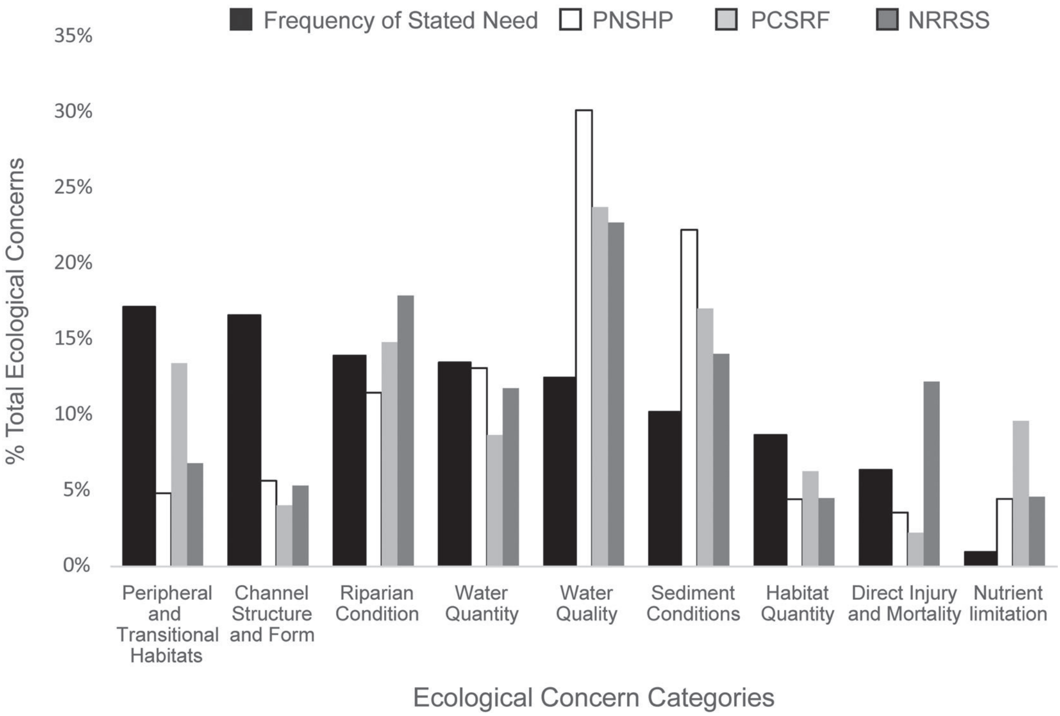

Fig. 2. Comparative histogram illustrating the relationship between the choice of ontology and inferences on how the distribution of projects matches the environmental concerns expressed for the Columbia River Basin. This figure is equivalent to Fig. 3A in Barnas et al. (2015), but with the addition of the distribution of habitat restoration actions when defined in the alternative ontologies. The black bars are the relative frequency that the labeled ecological concern is expressed in the sub-basin plans across the Columbia River Basin. The white, gray, and dark gray bars are the distribution of projects when defined by the Pacific Northwest Salmon Habitat Project (PNSHP), Pacific Coast Salmon Recovery Fund (PCSRF), and National River Restoration Science Synthesis (NRRSS) ontologies, respectively.

respectively. While there was variety in the correlations, with the PNSHP and NRRSS data systems being most similar, none were statistically significant.

Discussion

Our results confirm that representing the same data with three different ontologies alters the information conveyed by that representation, and thus a researcher's likely interpretation. Project type distributions varied, and their ability to express relationships was dependent on the choice of ontology. In general, the PNSHP and PCSRF distributions were most similar in terms of rank order of project frequency (Fig. 1), but the PNSHP and NRRSS representations were more similar in expressing underlying processes (Fig. 2). These results suggest that recognition of uncertainty introduced by the design of the data system needs to be a consideration in any data federation effort, but in evaluating this effect it is useful to know how big the effect is and what some of the potential consequences of that effect are.

While we have reported statistical differences in various correlation estimates, it is not obvious if these differences have real-world significance, nor is it clear that differences in correlation coefficient here would be reflected more generally in other types of data. In this case study however, the effect of using different data dictionaries to represent the same data has the potential to change qualitative answers about restoration priorities. For example, were one to switch from using the PNSHP to the NRRSS data definitions to represent the same data, water quality improvement projects would go from one of the rarest and most expensive restoration actions undertaken to the most common and least expensive type of management actions. The reason for this discrepancy lies in the motivations and logic for the different data synthesis projects rather than data quality. The NRRSS project was designed to address ecological process, and so, restoration actions that altered sediment flow or water temperature resulting in improved water condition were classified as water quality projects. In general, these types of water quality projects were numerous and cheap. Conversely, the PNSHP data definition captured the physical actions taken by the restoration project sponsor. So, water quality projects, such as toxic waste or mine tailing cleanups, are fewer in number and happen to be very expensive.

While the reasons behind these findings may be rational and intentioned, the consequences have high stakes. In particular, the ontology chosen has implications for evaluation of implementation and monitoring of habitat restoration in the Pacific Northwest. In a national synthesis of habitat restoration, more than three quarters of all actions in the United States were in the Pacific Northwest (Bernhardt et al. 2005). In the Columbia River Basin alone, habitat management expenditures have been estimated at $300–400 million per year in federal dollars (G.A.O. 2002), and more recent estimates indicate that this investment has continued in the years since and may exceed half a billion dollars per year when surrounding areas are included (S. L. Katz et al., *unpublished manuscript*). These costs do not include the ongoing litigation surrounding the regulatory and management context in the Columbia River Basin, within which environmental mitigation via restoration is a major component (Kareiva et al. 2000, G.A.O. 2002). In sum, the habitat restoration enterprise is an important economic engine in the region, and these data systems are the mechanism for maintaining accountability of, and compliance monitoring for the public money spent on the actions. It is important, therefore, that the broader community understand the sources of uncertainty that exist within these federated data systems. If a party had an interest in highlighting one type of habitat problem, water quality in the example above perhaps, it is important for all parties to understand the role of the data structures in biasing the representation of that information one way or the other.

Is there a basis for choosing one ontology over another, or identifying which one is best? As mentioned above, in assembling a data system, the intent is for the simplified representation (i.e., the data) to include enough of the available and relevant information from the real world to capture processes of interest and test models of those processes. The more accurate our representations, the more we will capture the signal of the underlying process with less noise from alternative processes. It is seductive to suggest that the best ontology is the one that has the highest signal-to-noise ratio. When we asked how restoration type frequency related to cost, the strongest signal was demonstrated by the PNSHP ontology, indicated by the highest magnitude correlation coefficient. This might lead us to suggest the PNSHP ontology is the best one. As defined however, this results in part from the internal logic of this ontology—to look at what action was completed. It is possible that if we had posed questions focused on constituencies, the PCSRF ontology may have had a better signal-to-noise ratio. It is also important to recognize that the developers of all of these data systems report that the choice of ontology was satisfactory for their needs (Bernhardt et al. 2007). Thus, we have a reasoned basis for suggesting that signal-to-noise ratio is a method for choosing one ontology among alternatives, but at the same time we must remember that because the ontology is also a model of nature (Guarino 1995), its performance can be conditioned on the context for the original data system design.

How much must humans be involved in data federation?

The challenge to interoperability produced by heterogeneity among candidate federated data systems is a widely recognized problem (Litwin et al. 1990, Qian 1993, Hammer and McLeod 1999, Haas et al. 2002, Jones et al. 2006, Madin et al. 2007). Among these studies, differences exist even in how heterogeneity itself is characterized. For example, Qian (1993) dichotomizes semantic mismatch, arising from contextual differences among data management entities, from representational mismatch, arising from different mappings of data language onto the information represented in a data system. Hammer and McLeod (1999) distinguish domain mismatch, which is similar to Qian's semantic mismatch, from schema mismatch, which can arise from topological differences among candidate data systems. In the ETL literature, the hierarchies of heterogeneity can be even more complex (Vassiliadis and Simitsis 2009, Theodorou et al. 2014). In addition, heterogeneity among source data can exist by degree rather than being the same or different. For example, in describing an IBM data synthesis product DB2 for example, Haas et al. (2002) characterize a gradient from tightly coupled data (sensu Litwin et al. 1990) with little heterogeneity and easy ingestion into federated systems, to efforts to combine data sets of unknown heterogeneity that may require custom applications, or wrappers, to resolve. Similarly, Madin et al. (2007), in describing an approach to using the open-source ontology development product OBOE, organize data with a framework —based on classes of data and the relationships among them—that mitigates semantic heterogeneity in the data. However, residual heterogeneity in these framework ontologies that hinders further interoperability may still require resolution with the use of semantic annotation provided by input from domain expertise (Madin et al. 2007).

Across the larger discussion of data federation exists an underlying expectation that there is some degree of common core semantic agreement among heterogeneous systems. That common core could be topological, some natural language overlap, or an explicit relational database, but the larger this common core, the less input from users or domain expertise is anticipated. As the scientific enterprise shifts further toward web or cloud basis, the pressure for data federation processes to be more automated and faster will also increase. This pressure includes both calls for the propagation of a single, or small number of similar data standards on the one hand (Katz et al. 2007, Madin et al. 2008, Horsburgh et al. 2011, Tarboton et al. 2011), and increasing automation in data discovery, ingestion, and synthesis on the other (Rahm and Bernstein 2001, Doan et al. 2004, Noy 2004). Environmental management in the Pacific Northwest is also on this trajectory to higher velocity data. Data standards would be most useful however, if those standards can be established and propagated prior to data collection. Unfortunately, in the case of habitat restoration, similar to other natural resource management problems, we are confronted by a 20-year legacy of diverse data collected without standards (Katz et al. 2007), and the point where we can deploy automation is still some distance in the future (Doan et al. 2004).

Currently, we are operating in a space where some degree of domain expertise is still required to resolve overlapping data dictionaries or ontologies. Apart from the potentially large effort and expense accessing subject matter expertise in the federation enterprise (Haas et al. 2002), we have shown here that there are potentially important impacts on the inferences drawn with that data. Specifically, we would argue that subject matter expertise is indispensable in evaluating interpretations based on federated data, and that significant consideration of this effect is appropriate when evaluating choice of ontology where legacy data are being combined. In our case, we are able to screen for these differences with estimates of correlations and signal to noise in test-bed inferences. Although the specific tests will vary in other applications, some characterization of data system behavior seems prudent (Hull 1997). Important epistemic differences emerge from the context, perspectives, and research questions of the parties collecting the original data leading us to recommend that some validation accompany the reporting of data federation projects in the future.

Acknowledgments

This work was funded in part by NSF EAR #1639458 and USDA #2017-67004-26131 awards, and the ongoing maintenance of the data system in this case study has been funded by National Oceanographic and Atmospheric Administration's National Marine Fisheries Service. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or policies of the National Science Foundation or the National Oceanographic and Atmospheric Administration. The authors are unaware of any outstanding conflicts of interest in the findings presented here. The authors thank A. Cramer, B. Fiest,

C. Jordan, and reviewers for helpful comments on earlier drafts.

Literature Cited

Barnas, K., and S. L. Katz. 2010. The challenges of tracking habitat restoration at various spatial scales. Fisheries 35:232–241.

Barnas, K. A., S. L. Katz, D. E. Hamm, M. C. Diaz, and C. E. Jordan. 2015. Is habitat restoration targeting relevant ecological needs for endangered species? Using Pacific salmon as a case study. Ecosphere 6:1–42.

Bayer, J. M. 2006. The Pacific Northwest Aquatic Monitoring Partnership: a forum for regional coordination. American Fisheries Society Symposium. 235. American Fisheries Society, Bethesda, Maryland, USA.

Bernhardt, E. S., M. A. Palmer, and J. D. Allan. 2005. Restoration of US rivers: a national synthesis. Science 308:636–637.

Bernhardt, E. S., E. B. Sudduth, and M. A. Palmer. 2007. Restoring rivers one reach at a time: results from a survey of US river restoration practitioners. Restoration Ecology 15:482–493.

Brunt, J. W., and W. K. Michener. 2009. The resource discovery initiative for field stations: enhancing data management at North American biological field stations. BioScience 59:482–487.

Carpenter, S. R., E. V. Armbrust, and P. W. Arzberger. 2009. Accelerate synthesis in ecology and environmental sciences. BioScience 59:699–701.

Chandrasekaran, B., J. R. Josephson, and V. R. Benjamins. 1999. What are ontologies, and why do we need them? IEEE Intelligent Systems and Their Applications 14:20–26.

Doan, A., J. Madhavan, P. Domingos, and A. Halevy. 2004. Ontology matching: a machine learning approach. Pages 385–403 *in* Handbook on ontologies. Springer, New York, New York, USA.

Dong, X. L., and F. Naumann. 2009. Data fusion: resolving data conflicts for integration. Proceedings of the VLDB Endowment 2:1654–1655.

G.A.O. 2002. Columbia River Basin salmon and steelhead: federal agencies' recovery responsibilities, expenditures and actions. GAO-02-612. U.S. General Accounting Office, Washington, D.C., USA.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5:199–220.

Guarino, N. 1995. Formal ontology, conceptual analysis and knowledge representation. International Journal of Human-Computer Studies 43:625–640.

Haas, L. M., E. T. Lin, and M. A. Roth. 2002. Data integration through database federation. IBM Systems Journal 41:578–596.

Hamm, D. E. 2012. Development and evaluation of a data dictionary to standardize salmonid habitat assessments in the Pacific Northwest. Fisheries 37:6–18.

Hammer, J., and D. McLeod. 1999. Resolution of representational diversity in multidatabase systems. Pages 91–117 *in* A. Elmagarmid, M. Rusinkiewicz, and A. Sheth, editors. Management of heterogeneous and autonomous database systems. Volume 4. Morgan Kaufman Publishers, San Francisco, California, USA.

Heidorn, P. B. 2008. Shedding light on the dark data in the long tail of science. Library Trends 57:280–299.

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky. 2011. Components of an environmental observatory information system. Computers & Geosciences 37:207–218.

Hull, R. 1997. Managing semantic heterogeneity in databases: a theoretical prospective. Pages 51–61 *in* Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. ACM, New York, New York, USA.

Jones, M. B., M. P. Schildhauer, O. J. Reichman, and S. Bowers. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37:519–544.

Kareiva, P., M. Marvier, and M. McClure. 2000. Recovery and management options for spring/summer chinook salmon in the Columbia River Basin. Science 290:977–979.

Katz, S. L., K. Barnas, R. Hicks, J. Cowen, and R. Jenkinson. 2007. Freshwater habitat restoration actions in the Pacific Northwest: a decade's investment in habitat improvement. Restoration Ecology 15:494–505.

Kolb, T. L., E. A. Blukacz-Richards, and A. M. Muir. 2013. How to manage data to enhance their potential for synthesis, preservation, sharing, and reuse—a Great Lakes case study. Fisheries 38:52–64.

Kuhn, W. 2003. Semantic reference systems. International Journal of Geographical Information Science 17:405–409.

Litwin, W., L. Mark, and N. Roussopoulos. 1990. Interoperability of multiple autonomous databases. ACM Computer Survey (CSUR) 22:267–293.

Madin, J. S., S. Bowers, M. P. Schildhauer, and M. B. Jones. 2008. Advancing ecological research with ontologies. Trends in Ecology & Evolution 23:159–168.

Madin, J., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2:279–296.

Noy, N. F. 2004. Semantic integration: a survey of ontology-based approaches. ACM SIGMOD Record 33:65–70.

Obrst, L. 2003. Ontologies for semantically interoperable systems. Pages 366–369 *in* Proceedings of the Twelfth International Conference on Information and Knowledge Management. ACM, New York, New York, USA.

Pierre, M. S., and W. P. LaPlant. 2000. Issues in crosswalking content metadata standards. NISO, Baltimore, Maryland, USA.

Qian, X. 1993. Semantic interoperation via intelligent mediation. Pages 228–231 *in* Proceedings RIDE-IMS'93: Third International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems. IEEE, New York, New York, USA.

Quine, W. V. 1970. On the reasons for indeterminacy of translation. Journal of Philosophy 67:178–183.

R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahm, E., and P. A. Bernstein. 2001. A survey of approaches to automatic schema matching. VLDB Journal 10:334–350.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.

Romanello, S., J. Beach, S. Bowers, M. Jones, B. Ludäscher, W. Michener, D. Pennington, A. Rajasekar, and M. Schildhauer. 2005. Pages 28–32 *in* Creating and providing data management services for the biological and ecological sciences: science environment for ecological knowledge. SSDBM, Santa Barbara, USA.

Schick, K. 1973. Indeterminacy of translation. Journal of Philosophy 69:818–832.

Shu, N. C., B. C. Housel, R. W. Taylor, S. P. Ghosh, and V. Y. Lum. 1977. EXPRESS: a data extraction, processing, and restructuring system. ACM Transactions on Database Systems 2:134–174.

Sokal, R. R., and F. J. Rohlf. 1969. The principles and practice of statistics in biological research. WH Freeman, San Francisco, California, USA.

Tarboton, D. G., et al. 2011. Data interoperability in the hydrologic sciences. Environmental Information Management Conference, Santa Barbara, California, USA.

Theodorou, V., A. Abelló, and W. Lehner. 2014. Quality measures for ETL processes. Pages 9–22 *in* International Conference on Data Warehousing and Knowledge Discovery. Springer, New York, New York, USA.

Uschold, M., and M. Gruninger. 2004. Ontologies and semantics for seamless connectivity. ACM SIGMOD Record 33:58–64.

Vassiliadis, P., and A. Simitsis. 2009. Extraction, transformation, and loading. Pages 1095–1101 *in* Encyclopedia of database systems. Springer, New York, New York, USA.

Volk, C. J., Y. Lucero, and K. Barnas. 2014. Why is data sharing in collaborative natural resource efforts so hard and what can we do to improve it? Environmental Management 53:883–893.

Williams, J., and S. G. Labou. 2017. A database of georeferenced nutrient chemistry data for mountain lakes of the Western United States. Scientific Data 4:170069.

Wright, C. 2017. Indeterminacy of translation. Pages 670–702 *in* Companion to the Philosophy of Language. John Wiley & Songs, Chichester, UK.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.2920/full