CDI FY17 Request for Proposals

Developing Best Practices for the Collection, Archival and Release of Genetic Data Across USGS

Submission Title: Developing Best Practices for the Collection, Archival and Release of Genetic Data Across USGS

Lead PI: John Pearce

Mission Area: Ecosystems

Region: Alaska

Organization: Alaska Science Center

OrcId: 0000-0002-8503-5485

Phone: 9077867094

Email: jpearce@usgs.gov

City: Anchorage

State: AK

Co-PIs and Collaborators:

Type: CO-PI

Name: Sandra Talbot

Mission Area: Ecosystems

Region: Alaska

Organization: Alaska Science Center

OrcId: 0000-0002-3312-7214

Phone: 9077867188

Email: stalbot@usgs.gov

City: Anchorage

State: AK

Type: CO-PI

Name: Barbara Pierson

Mission Area: Ecosystems

Region: Alaska

Organization: Alaska Science Center

OrcId: 0000-0001-8233-874X

Phone: 9077867000

Email: bpierson@usgs.gov

City: Anchorage

State: Alaska

Science Support Framework Element 1: Communities of Practice

Science Support Framework Element 2: Data Management

Science Support Framework Element 3: Science Data Lifecycle - Publishing/Sharing

In-Kind Match: $55,359.00

List of anticipated deliverables from the project: 1. Survey USGS Science Centers to assess degree of proficiency with metadata publication. 2. Establish a genetic data “forum” or internal website for users to facilitate document/template exchange. 3. Coordinate sharing of current data dictionaries, metadata software interfaces, process steps, data quality procedures, areas of greatest concern and evaluate for strengths, weaknesses, best options. 4. Create documentation outlining effective metadata generation to coincide with actual project timelines. 5. Streamline the generation of metadata by standardizing and populating background data to be included in metadata (e.g., location and date of collection, age, sex, etc.) upon receipt of samples. 6. Create data dictionaries, website lists with genetics specific keyword guidance, spreadsheet templates,

decision trees for whether to include certain types of data. 7. A summary of the new standards, quality control measures, and data archival and release processes that detail and make public the processes used by USGS genetics laboratories across the country, and be citable by USGS users.

Lead Cost Center: USGS Alaska Science Center

Notes, Comments:

Project Description: Genetic data are one of the most widely shared information sources in the biological sciences. Nucleotide data have been accessioned into GenBank or Dryad and queried for years. Most peer-reviewed scientific journals have required data release via such databases prior to publication. However, neither GenBank nor Dryad satisfy the FGDC Standard requirement. The Alaska Science Center has released genetic data on its Data Release page (http://alaska.usgs.gov/products/data_all.php) in compliance with the FGDC standard. In an examination of ASC released data set website statistics, 5 genetic data sets were among the top 20 web pages viewed. Unique page views ranged from 41 to 223 and average time spent on pages ranged from 1 to 3 mins. Given the new data policies, limited coordination among USGS molecular genetic laboratories, and apparent interest in USGS genetic data, we propose to facilitate greater collaboration among USGS geneticist (already contacted about this proposal) from the ASC, Fort Collins Science Center, Leetown Science Center, Forest and Rangeland Ecosystem Science Center, and the Western Ecological Science Center. Funding will support half of a term salary to develop, coordinate, and publish common standards for genetic data collection, minimum quality control measures, and data and metadata description and release methods so that USGS genetic data are being released in a highly similar format and quality across the country.

Total Budget: $49,610.00

SECTION 1. PROJECT SUMMARY

Title: Developing Best Practices for the Collection, Archival and Release of Genetic Data Across USGS

Name: John Pearce

Genetic data are one of the most widely shared information sources in the biological sciences. Nucleotide data have been accessioned into GenBank or Dryad and queried for years. Most peer-reviewed scientific journals have required data release via such databases prior to publication. However, neither GenBank nor Dryad satisfy the FGDC Standard requirement. The Alaska Science Center has released genetic data on its Data Release page (http://alaska.usgs.gov/products/data_all.php) in compliance with the FGDC standard. In an examination of ASC released data set website statistics, 5 genetic data sets were among the top 20 web pages viewed. Unique page views ranged from 41 to 223 and average time spent on pages ranged from 1 to 3 mins. Given the new data policies, limited coordination among USGS molecular genetic laboratories, and apparent interest in USGS genetic data, we propose to facilitate greater collaboration among USGS geneticist (already contacted about this proposal) from the ASC, Fort Collins Science Center, Leetown Science Center, Forest and Rangeland Ecosystem Science Center, and the Western Ecological Science Center. Funding will support half of a term salary to develop, coordinate, and publish common standards for genetic data collection, minimum quality control measures, and data and metadata description and release methods so that USGS genetic data are being released in a highly similar format and quality across the country.

Goals in the Context of Evaluation Criteria:

This proposal focuses on a targeted effort requiring coordination across a number of Centers, Regions, and types of funding provided to these groups. This effort leverages existing capabilities at each Center and those of the Data Management Group at the ASC, which has developed standards for the archival and release other types of information that can be applied to USGS genetic data. This effort will provide guidance to genetics researchers on how to best preserve, display and improve access to the USGS genetic information through the development of best practices for these kinds of data. The effort will also help avoid duplication of effort. Lastly, this project will touch on every stage of the data lifecycle from planning to data release.

Timeline and Deliverables:

1. Survey current centers to assess degree of proficiency with metadata publication.
2. Set up a genetic data “forum” or internal website for users to facilitate document/template exchange.
3. Coordinate sharing of current data dictionaries, metadata software interfaces, process steps, data quality procedures, areas of greatest concern and evaluate for strengths, weaknesses, best options.
4. Create documentation outlining effective metadata generation to coincide with actual project timelines.
5. Streamline the generation of metadata by standardizing and populating background data to be included in metadata (e.g., location and date of collection, age, sex, etc.) upon receipt of samples.
6. Create data dictionaries, website lists with genetics specific keyword guidance, spreadsheet templates, decision trees for whether to include certain types of data.
7. A summary of the new standards, quality control measures, and data archival and release processes that detail and make public the processes used by USGS genetics laboratories across the country, and be citable by USGS users.

SECTION 2. ESTIMATED BUDGET

Budget Category	Federal Funding "Requested"	Matching Funds "Proposed" Paid by Alaska Science Center
1. PERSONNEL (SALARIES including benefits):		
Federal Personnel Total:	$39,500	$45,752
Contract/Collaborator Personnel Total:		
Total Salaries:	$39,500	$45,752
2. TRAVEL EXPENSES:		
Travel total	$1,500	
Other Expenses		
Total Travel Expenses:	$1,500	$0.0
3. OTHER DIRECT COSTS:		
Equipment:		
Publication Costs:		
Office Supplies, etc.		
Total Other Direct Costs:	$0.0	$0.0
Total Direct Costs:	$41,000	$45,752
Indirect Costs (21%):	$8,610	$9,607
GRAND TOTAL:	$49,610	$55,359