Data: who has it, where is it, and how to get it

A nested case study illustrates the challenges of cataloguing data related to water resources, and finding and obtaining data in the Coeur d'Alene Basin, Idaho

Karen Trebitz, PhD student, Water Resources, University of Idaho, Moscow; 2017

Section I. The Challenge

"Identify individuals, governmental agencies (state, federal, Tribal) and any NGOs responsible for, or associated with the collection of water quality data … across the Columbia River Basin (rivers and lakes – from headwaters to Bonneville Dam)…" (from narrative of assignment).

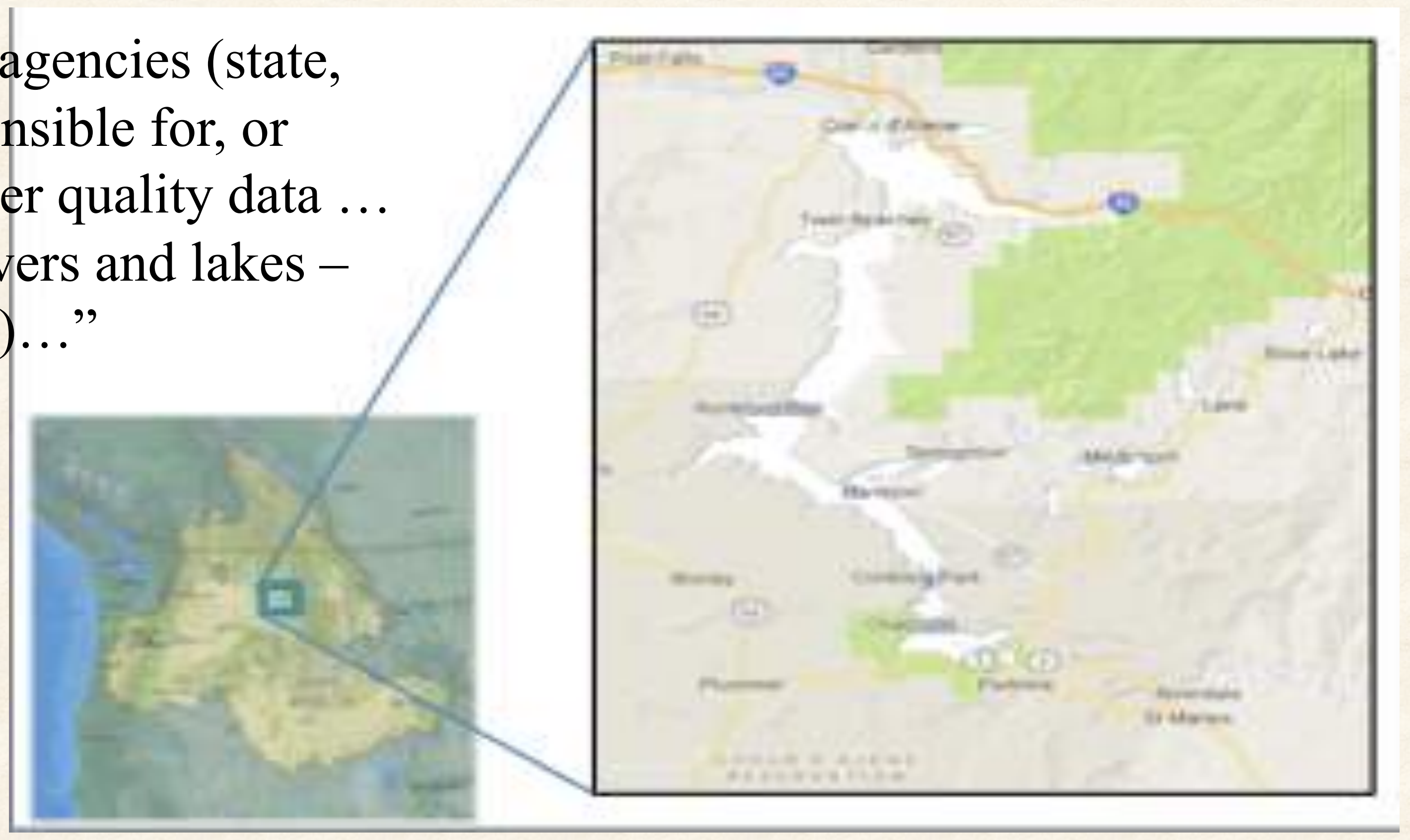


Figure 1. The Columbia River Basin (left; Coeur d'Alene Subbasin (right)

Target area for study

The Coeur d'Alene Subbasin of the greater Columbia River Basin, North Idaho (Figure 1)

Methods

Determine the criteria for boundaries

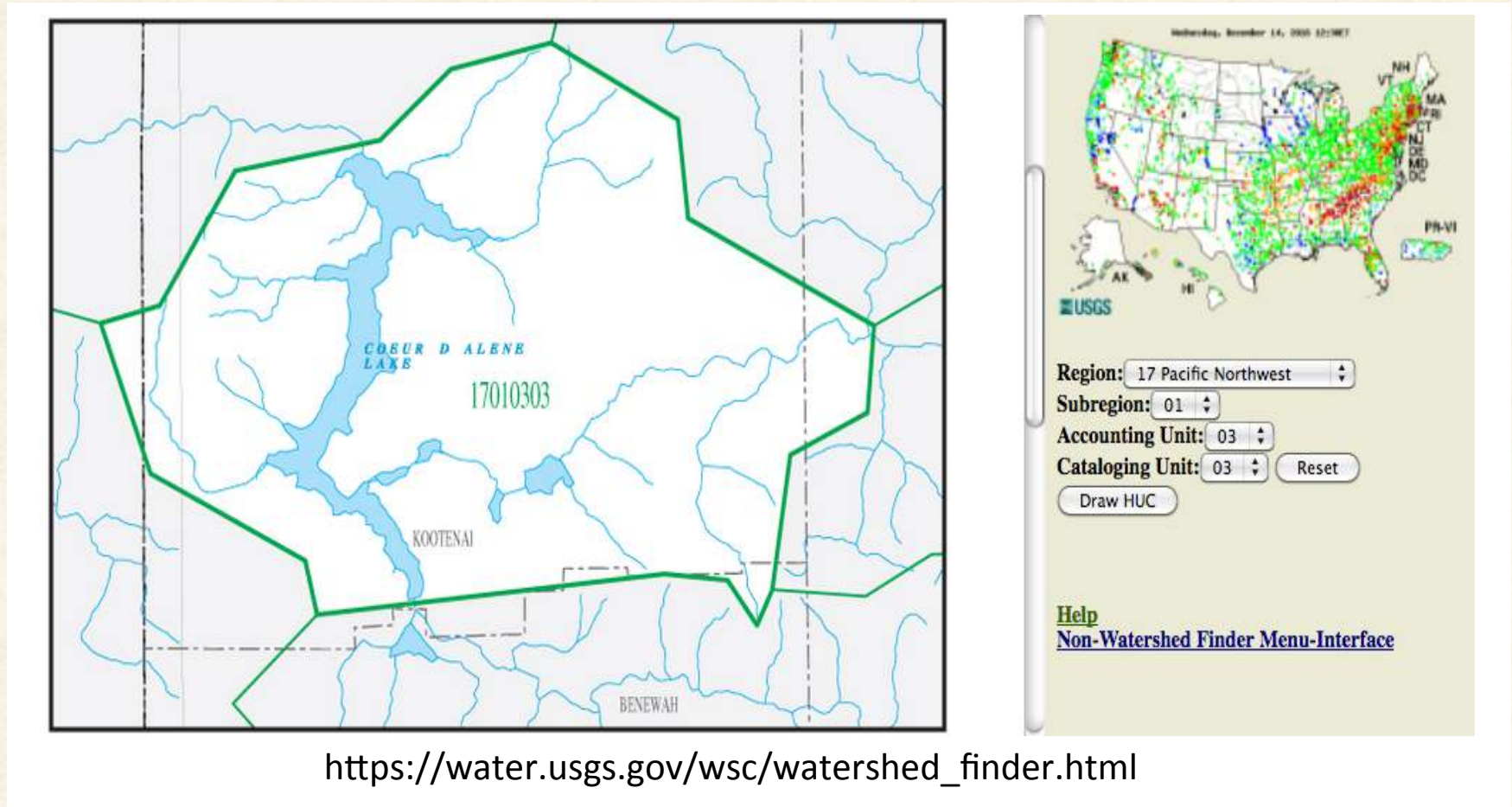
- Geographic delineation (Figure 2)
- Geopolitical boundaries (states, Tribal)

Define the criteria for water quality

- Primary definitions
- Secondary, related factors

Identify "actors" who might have data

- Explore actors' websites
    - Mention of and links to data;
    - Compile data matrices from interactive websites
    - Follow links to partner actors and other websites → iterative
- Contact by phone and/or email to
    - Get as much information as possible, and possibly a brief conference
    - Ask about other possible places and people; and ask for introductions → best results!
    - Accept any data offered
    - Offer to share useful information



https://water.usgs.gov/wsc/watershed_finder.html

Figure 2. Defining boundaries by HUC (Hydrologic Unit Code)

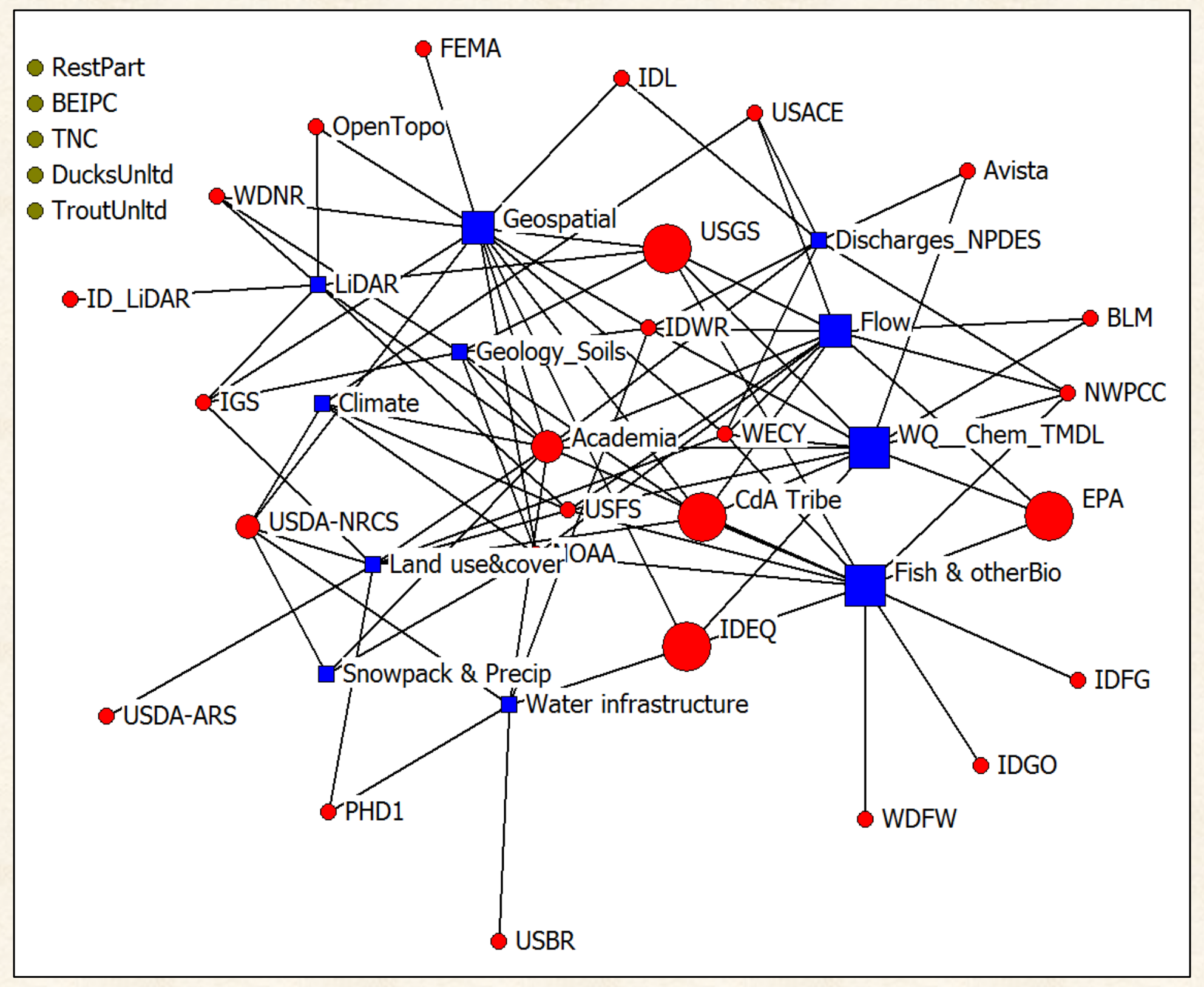


Figure 3. Network map of relationships of actors (red dots) to data types (blue squares); node size relates to size of data collections (actors) and the importance value given to the data types in water quality considerations.

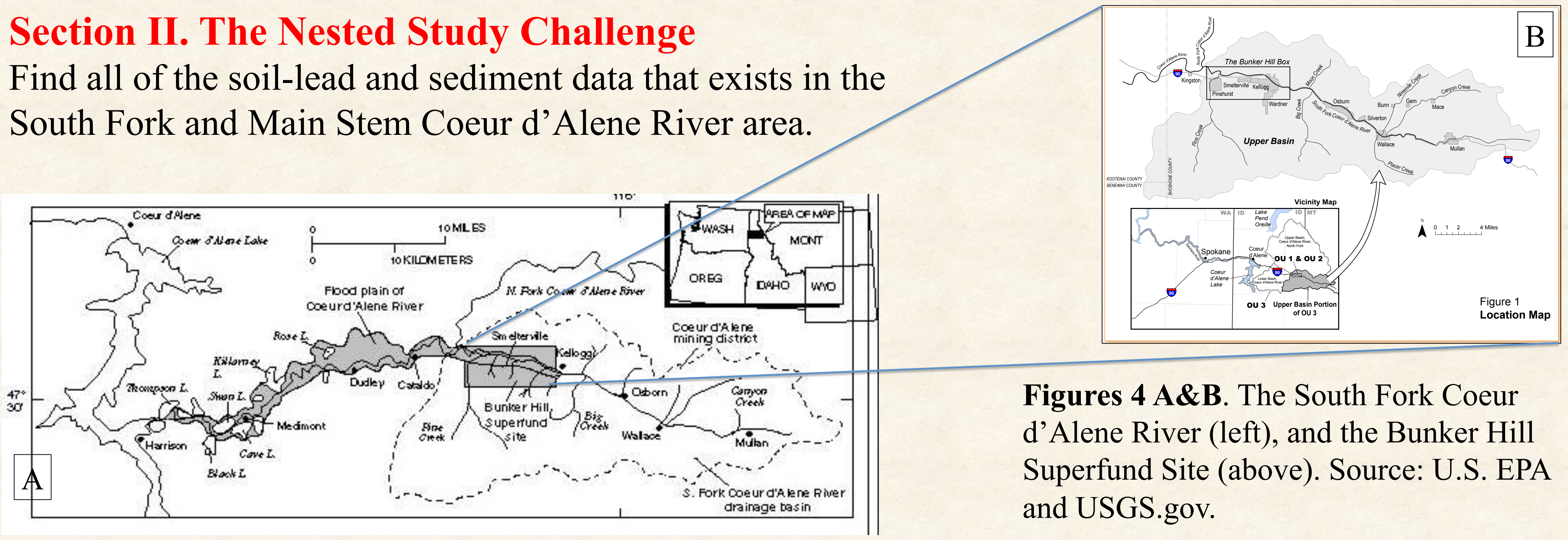
Product: Report to aid in data searches

- Overview of watershed
- Discussion of open access laws
- Summary of actors and data (network model, Figure 3)
- Matrices of some major datasets

Acknowledgements: To team members Alex Suchar and Roger Lew for their ongoing encouragement. *This research was supported by National Science foundation (NSF) award #1249400, and NSF award #IIA-1301792, Miles Track 1, from the NSF Idaho EPSCoR Program.* Report to be published by Idaho Water Resources Research Institute.

Section II. The Nested Study Challenge

Find all of the soil-lead and sediment data that exists in the South Fork and Main Stem Coeur d'Alene River area.



Figures 4 A&B. The South Fork Coeur d'Alene River (left), and the Bunker Hill Superfund Site (above). Source: U.S. EPA and USGS.gov.

Purpose

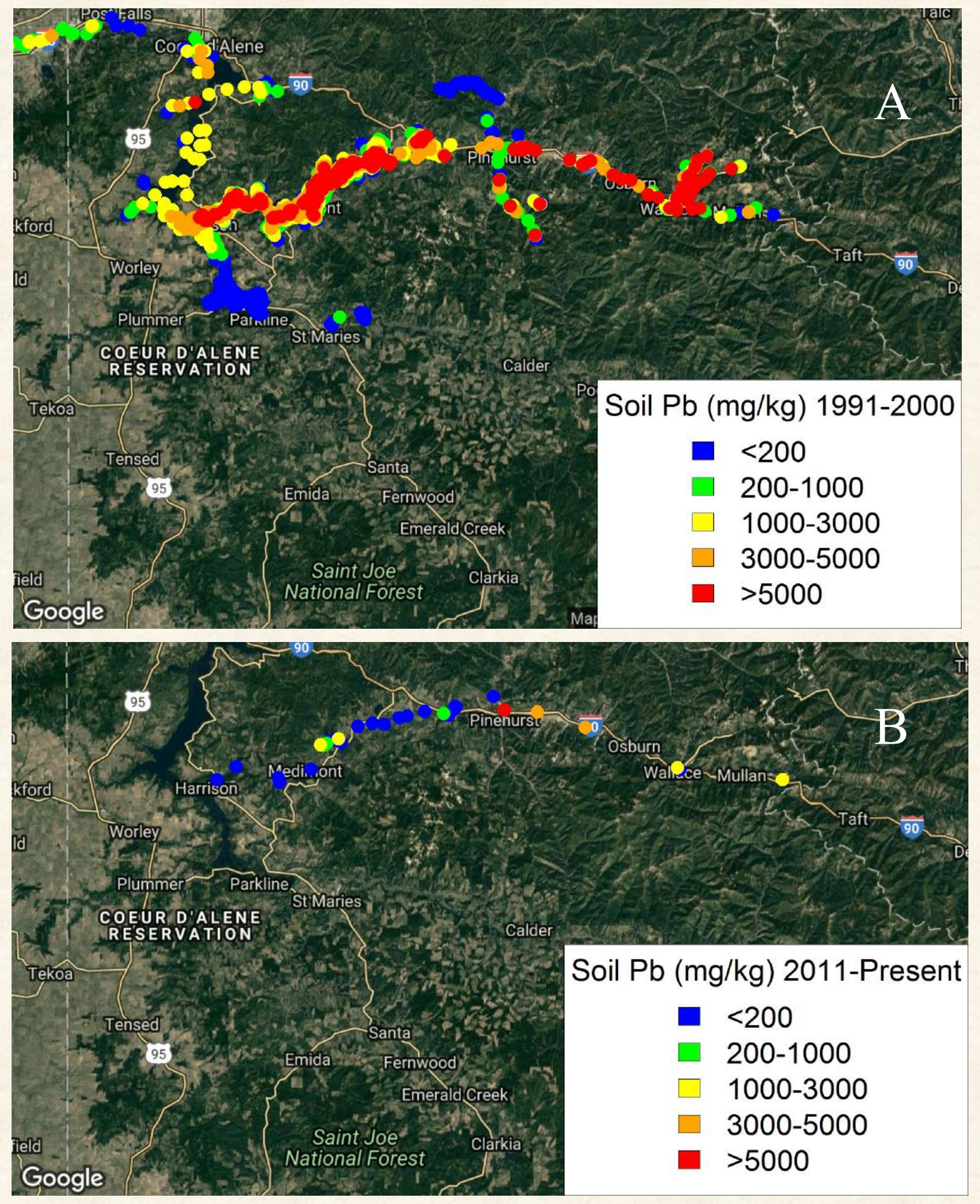
Background soil-lead levels are needed to develop an interactive, web-based model and game to increase public awareness of lead exposure risks when recreating in the river corridor. The Coeur d'Alene River is a popular recreation destination, but it is also home to the Bunker Hill Superfund Site (Figures 4 A& B). Legacy heavy metals contamination persists throughout the basin, including into Coeur d'Alene Lake and downstream to the Spokane River.

Methods

Boundaries and ***Data parameters*** are given

Actors

- Begin with the data summary from Section I to most efficiently find probable data sources.
- Use URLs and hyperlinks to access online information.
- Contact actors and for datasets
- Compile datasets into one single-format dataset



Figures 5 A&B. Lead concentrations in soils and sediments 1991-200 (top), and 2011-2016 (bottom). Source: Alex Suchar.

Results

Data were obtained through the contacts and web URLs covered in Section I. Perhaps 90 percent of the data were from three major studies that spanned almost 30 years. Data obtained from 4 different actors were variations of the same large dataset from 2001, but with transcription errors and pieces of information missing. Studies contracted by the EPA used fairly consistent lab protocols, but other data in this set could be questionable. The biggest difficulty was in at least 4 different formats of geolocations. We ended with approximately 1,500 usable data points; very few from the past decade (Figures 5 A&B). Values in Figure 5B are misleading, as many of these points are monitored, remediated sites that are cleaned annually.

Discussion and Conclusions

The information and network tools in the report from Section I were useful when applied to the specific data search in Section II. What was troublesome was the inordinate amount of time spent asking around for datasets that everyone "knew" existed, and were supposedly "well-documented." Datasets were not named logically, not curated with metadata, and the institutional memory of them is being lost with time.