1.0 Project Summary

Name of project: Modeling data and information needs for avian conservation using Neo4j

Project lead and contact details:
Brian Wee, Massive Connections LLC, bwee@massiveconnections.com, ORCID: 0000-0002-0038-9381

Project partners and contact details:
Jessica L. Burnett, U.S. Geological Survey, jburnett@usgs.gov, ORCID: 0000-0002-0896-5099
Steven Aulenbach, U.S. Geological Survey, saulenbach@usgs.gov, ORCID: 0000-0002-0172-6538
William Teng, NASA GES DISC, william.l.teng@nasa.gov, ORCID: 0000-0001-5642-6359
Rustem A. Albayrak, NASA GES DISC, rustem.a.albayrak@nasa.gov, ORCID: 0000-0001-6060-9844

Proposed start and end date: August 1, 2020 to January 31, 2021

Budget Requested: Total of **$3388.80** (details of deliverables described in section 2.4)

- Deliverable #1: Conceptual Model of Data and Information for Conservation Decision Making (Report). **$1500.**
- Deliverable #2: Neo4j instance with prototype data (JSON export and documentation). **$1500.**
- Neo4j Aura (Neo4J cloud hosted Database as a Service): 6 months * $64.80 per month = **$388.80**.

2.0 Project Outline

2.1 Overview

We propose to assess the utility of a labeled property graph database to help elucidate how data might be applicable for a given avian conservation application. We have conducted a brief assessment of triple stores versus labeled property graphs, and have elected to experiment with the latter. Furthermore, we have elected to use Neo4j because NASA is a Neo4j customer and members of our team have been in contact with the company's customer support. Our graph will likely utilize a schema that reflects selected entities involved in avian conservation. An initial assessment suggests that minimally, modeled entities include: management goals, threats to endangered avian populations, federally determined lists of threatened avian species, and existing data products that estimate essential avian population demographics. A successful Neo4j query will show, for example, a series of connected node-relationship-node (i.e. subject-predicate-object) triples -- a traceability subgraph -- that reflect the process of how a data product was used to estimate avian population demographics for populations that are likely to be impacted by climate change. The traceability subgraph represents, in effect, a "data to decisions" pathway. If the proposed project succeeds in demonstrating this capability, we envision future capabilities (developed under separate funding) to visualize and explore alternate traceability subgraphs within a given graph-space to assess alternate ways to connect data to decisions. This ultimately translates to a much needed capability for decision making where the requisite data for informed decision making is not available, and surrogate data resembling the decision making context must be used instead.

2.2 Foundations for the Proposed Project

The proposed project is built on the following efforts which has yielded promising directions and expressions of interest from various ESIP constituents. We are confident that the proposed project is feasible, highly aligned with the ESIP organizational vision and mission, and comprised of team members who have had a consistent demonstrated record of delivering on project goals:

A. **2017 Data to Decisions Cluster.** Based on collaborations with the US Climate Resilience Toolkit championed by the ESIP Agriculture and Climate Cluster (ACC), PI Wee initiated the "Data to Decisions Cluster" (charter initiation date: 2017-09-22, charter sunset date: 2017-12-31) (https://tinyurl.com/d2dresilience) that sought to "*collaboratively build openly web-accessible concept maps for climate adaptation*".

B. **2018 ESIP Lab Proposal.** Based on ideas developed in the 2017 Data to Decisions Cluster, the ESIP lab proposal "*Enabling the encoding and visualization of provenance metadata for better discovery and understanding of climate resilience strategies for agriculture-related decision-making*" (effective dates: July 1, 2018 to January 31, 2019) aimed to assess "*the technologies required to implement a 'GitHub' for resilience planning*" (https://github.com/ESIPFed/d2dprovenance). Deliverables from the project included a report titled "D2dprov: Vision 2025 A transdisciplinary science, technology, and policy synthesis on data-driven, science-informed resilience planning for 2025 and beyond".

C. **2020 Agriculture and Climate Cluster's (ACC) Community concept mapping activity March 2020 through July 2020**. This experiment (https://tinyurl.com/cy2020-ag-climate), hosted by the Agriculture and Climate Cluster and based on ideas developed in the 2018 ESIP Lab Proposal, developed data-to-decisions concept maps based on three use-cases (Chesapeake Bay nutrient loading under future climate, wildfire mitigation, and wildfire response). This work, mainly led by Wee, Teng, and Albayrak, has resulted in outreach and collaborations with individuals from the ESIP Disaster Lifecycle, the ESIP Semantics Technology Committee, the ESIP Air Quality Cluster, the USGS, NASA Goddard, NOAA NWS, USDA Forest Service, the University of Florida, and the University of Texas at Austin. The ACC is also hosting an ESIP summer meeting session on concept mapping (July 2020), and co-organizing with the ESIP Semantics Technology Committee an introductory tutorial on ontologies (July 2020, an ESIP summer meeting pre-conference event).

2.3 Science, Technology, and Policy Context for Proposed Project

The collaborations and intellectual capital developed between 2017 and 2020 outlined above were founded on fostering transdisciplinary outcomes through methods that are aligned with science-informed, data-driven decision making/policy development. The triumvirate of science, technology, and policy aspects for the proposed project are briefly outlined below.

A. **Science.** Co-PI Burnett is currently assessing methods for assessing the goodness-of-fit between (a) algorithmic (statistical) methods for assessing time-series data that harbor signals that indicate a regime change (e.g. a sudden disruption in net primary productivity over a certain period of time), and (b) the actual time-series observation data that are currently available. Certain data may not be amenable to a given statistical analysis, so this goodness-of-fit assessment is important. These observation data are existing digital artifacts that have been collected using different protocols, have varying degrees of metadata integrity, have different levels of uncertainty, and may have inadequately documented statistical transformations applied to them. These challenges apply to datasets that Burnett will examine as part of the proposed project, including a number of different datasets used by the avian conservation community.

B. **Technology.** Wee, Teng, and Albayrak, in one of their 2018 ESIP lab project deliverables (see section 2.2 (B)) proposed using knowledge graphs to model the data-to-decisions "pathway". During the first half of CY2020, we developed a repository of concept maps that correspond to these data-to-decisions "pathways" using the freeware tool Cmap (https://cmap.ihmc.us/). We also deliberated using Neo4j to host transformed versions of those concept maps to enable query and visualization.

C. **Policy.** The Migratory Bird Treaty Act (16 U.S.C. 703 et seq.) requires the US Fish and Wildlife Service to identify avian fauna that may qualify for protection under the Endangered Species Act (16 USC 2912 Sec. 13). Various wildlife conservation NGOs thus pay attention to the data used to inform the status of avian species. Under the rubric of the Foundations for Evidence-based Policymaking Act (Public Law 115-435: which also includes Title II - OPEN Government Data Act), technologies that foster traceability between data and policy decisions are of relevance to the ESIP community.

2.4 Key Project Tasks and Deliverables

A. **Deliverable #1: Conceptual Model of Data and Information for Conservation Decision Making (Report).** Develop a conceptual model that captures selected entities involved in avian conservation. An initial assessment suggests that minimally, modeled entities should include management goals, threats to endangered avian populations, federally determined lists of threatened avian species, and existing data products that estimate essential avian population demographics. Similar to the work described in section 2.3 (B), we shall consult existing ontologies (including the BFO-compliant ENVO, IAO, and RO ontologies) to guide our model development. Work funded under an earlier ESIP Lab Proposal (see section 2.2 (B)) suggests two decision making ontologies (Kornyshova & Deneckère, 2012 and Locher & Costa, 2017) that will be considered for modeling management decisions.

B. **Deliverable #2: Neo4j instance with prototype data (JSON export and documentation).** Deliverable #1 will be used to prototype a Neo4j instance that reflects a small number of avian demographic data products that are used by the conservation community. Due to potential sensitivities, a best attempt will be made to reflect the structure (i.e. columns / fields) of those data products, but not necessarily the actual values. Figure 1 shows a highly abstracted example of what the graph model may look like (key-value pairs of nodes and relationships not shown). We shall create a small number of exemplar Neo4j queries that reflect the types of questions that can be asked of the database. Queries will be executed in the "Neo4j Bloom" Graph App, which facilitates near natural language search. The database can be queried to answer questions like "*show the types of data products that are useful for assessing the impacts of advancing spring seasons on avian populations and the corresponding management goals.*" Development on the Neo4j Aura cloud instance will facilitate access to the database by team members and invited reviewers. We shall also experiment with executing Neo4j queries from within Python. At the end of the project, the Neo4j Aura cloud instance will be deactivated, and its data will be exported to JSON and shared on GitHub.

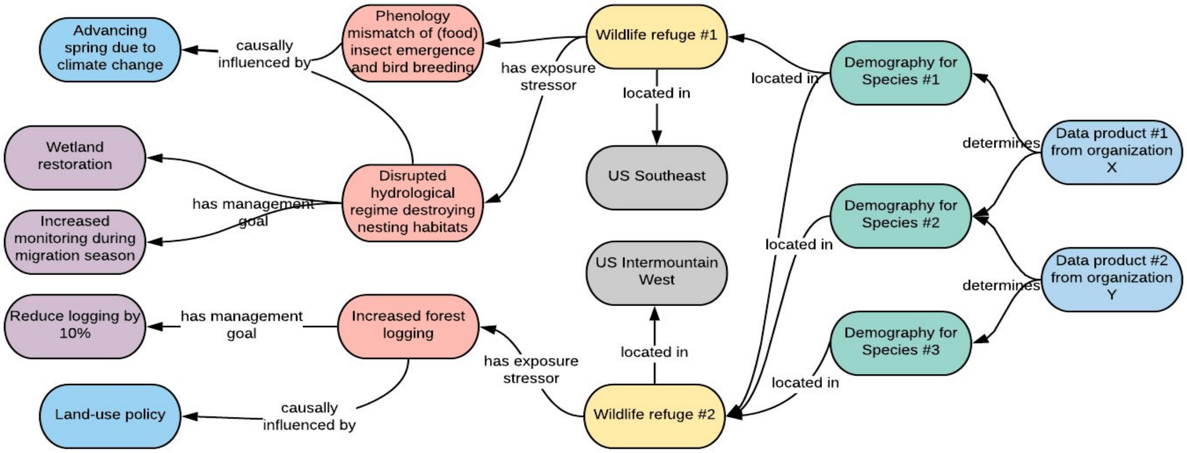
Advancing spring due to climate change
causally influenced by
Phenology mismatch of (food) insect emergence and bird breeding
has exposure stressor
Wildlife refuge #1
located in
US Southeast
located in
Demography for Species #1
determines
Data product #1 from organization X
Wetland restoration
has management goal
Increased monitoring during migration season
Disrupted hydrological regime destroying nesting habitats
located in
Demography for Species #2
US Intermountain West
determines
Data product #2 from organization Y
Reduce logging by 10%
has management goal
Increased forest logging
has exposure stressor
located in
Demography for Species #3
located in
Land-use policy
causally influenced by
Wildlife refuge #2

Figure 1: Highly simplified graph schema instantiated with exemplar data. All relations are terms defined in the OBO Foundry's Relations Ontology, except for "has management goal" which is based loosely on Kornyshova & Deneckère's 2012 Decision Making Ontology.

3.0 Outreach

Given the broad landscape of the proposed project covering avian management concerns, scientific subject matter, and technology matters, a carefully targeted engagement strategy is required. Attempting to cover the whole gamut of disciplinary topics with any one given audience will almost certainly end in failure. We anticipate an initial discussion with a small group of avian conservation SMEs to discuss conservation questions that benefit from data-driven approaches. We will then iterate further development of the conceptual model and Neo4j implementation before cycling back to the SMEs for further feedback. We anticipate reporting on our progress at the ESIP ACC monthly meetings as a continuation of the three data-to-decisions use-cases developed under the Cluster (see Section 2.2 (C)). These use-cases will be assessed for transformation and ingest into Neo4j. Project developments will be presented at ESIP meetings and other venues as appropriate. We intend to continue our pattern of extensive outreach (see Section 2.2 (C)) to ESIP constituents outside the ACC.

4.0 Project Partners

- Brian Wee (Massive Connections) | Principal Investigator (ecology, informatics, program management, stakeholder engagement, policy)
- Jessica L. Burnett (USGS) | Co-I (ecology, avian conservation, stakeholder engagement, stakeholder communications)
- Steven Aulenbach (USGS) | Co-I (provenance, discovery, data products)
- William Teng (NASA GES DISC) | Co-I (data products, informatics, program management, stakeholder engagement)
- Rustem A. Albayrak (NASA GES DISC) | Co-I (machine learning, data science, statistics, natural language processing)