The DataBridge: A Social Network for Long Tail Science Data

Howard Lander
howard@renci.org
Renaissance Computing Institute
The University of North Carolina at Chapel Hill
The DataBridge: A Social Network for Data
Dark Data from The Long Tail of Science

- Long tail data is the small data sets produced by numerous investigators.
- From Brahe to Mendel discovery has come from relatively small data sets.
- Much long tail data is dark data, data “not easily found by potential users” (Heidorn).
- Long tail data sets lack structural advantages of “classic” Big Data.
The DataBridge Strategy: Building a Social Network for Scientific Data

• Construct a multi-dimensional sociometric network for data. Three challenges:
  - Evaluate the similarity/relevancy of data sets
  - Perform community detection on the resulting set of similarities
  - Provide query interfaces on resulting multi-dimensional network
DataBridge Implementation

Visualization and Query Clients

Relevance Engines Including Relevance Algorithms and SNA

Network Database

AMQP Based Message Oriented Middleware

Metadata & Ontology Database

Data Reformatters

Data and Meta Data Gatherers

DataVerse Instances

iRODS Data Grids
DataBridge Progress to Date: JavaScript based network visualization tool

Visualization of data relationship Networks for DataBridge

This network shows the due-to-data relationship in Harris surveys extracted from Odum Institute DATaverse Network at UNC Chapel Hill. A categorical data similarity measurement algorithm was used to extract a similarity adjacency matrix that was then used to create this data-to-data relationship graph. Each node represents a Harris survey data room; each edge links the pair of nodes based on their similarity measurement — the thicker the edge, the more similar the linked nodes.
Visualization of data relationship Networks for **DataBridge**

This network shows the data-to-data relationship in Harris surveys extracted from [Odum Institute Dataverse Network](https://doi.org/10.1002/9781597451205) at UNC-Chapel Hill. A categorical data similarity measurement algorithm was used to extract a similarity adjacency matrix that was then used to create this data-to-data relationship graph. Each node represents a Harris survey data record; each edge links the pair of nodes based on their similarity measurement — the thicker the edge, the more similar the linked nodes.

Similarity measure between node Harris 1972 presidential election survey, no. 2240 and node Harris 1986 Black Women Leadership Survey, Study no. 864010 is 0.625
DataBridge Team

- **PI:** Arcot Rajasekar RENCI and SILS, UNC-Chapel Hill
- **Collaborators:**
  - Odum Institute, UNC-Chapel Hill
  - Population Informatics Research Group, UNC-Chapel Hill, Texas A & M University
  - iLab, North Carolina A&T University
  - The Institute for Quantitative Social Science, Harvard University
- **Funded by:** NSF Office of Cyberinfrastructure Awards OCI-1247562, OCI-1247602 and OCI-1247663