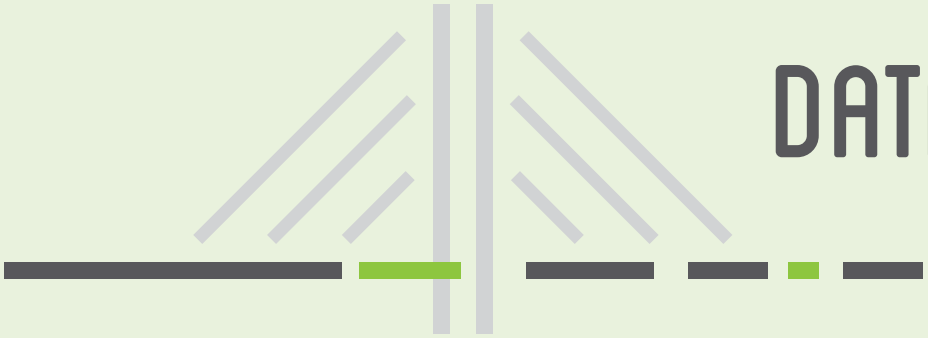
renci



DATABRIDGE

Vol. 3, No. 5 | July 2015
RENCI WHITE PAPER SERIES

DataBridge:

CREATING BRIDGES TO FIND DARK DATA

The Team

HOWARD LANDER
Senior Research Software Developer (RENCI)

ARCOT RAJASEKAR, PhD
Chief Domain Scientist, Data Grid Technologies (RENCI) and Professor, School of Information and Library Science, UNC-Chapel Hill

Additional writing support from Anne Frances Johnson and Kathleen Pierce.

Summary

Getting the most out of our research investments requires access to the full scope of data being collected. Yet all too often, researchers individually gather data, publish their analyses, and then store the data away, never to be used again. This "dark data" holds tremendous potential, but it is currently untapped because it simply isn't discoverable by others. Science needs new tools to find new connections among the millions of hidden scientific data sets and unlock the value within.

DataBridge is a streamlined, automated system that enables these valuable, existing data sets to be discovered, shared, and used. By creating a sophisticated social network that enables data sets to connect with each other, DataBridge aims to shine a light on dark data.

With DataBridge, researchers can discover existing data sets that are relevant to their research, find other scientists engaged in similar research, and contribute their data for the benefit of others. It illuminates hidden connections across the scientific world, allowing researchers to make unexpected discoveries from unexpected data--and allowing the country to reap more benefit from each dollar of its research investment.

The Challenge:
Finding Data

The scientific enterprise runs on the collection, analysis, and interpretation of data, both large data sets and small data sets. There are large data sets being generated in the sciences by NASA, the Large Hadron Collider, and many other government entities and research institutes. There are also large data sets being generated for business purposes, such as those being created and used by Facebook, Twitter, and Amazon.com. The first type of "big data" receives heavy media attention and multimillion-dollar grants. The second type is being optimized by companies for financial gain.

But there is also a third type of big data, whose potential is currently untapped. This is the universe of valuable small data sets created by individual researchers or research teams. Such data, sometimes knows as the "long tail of science" (Palmer 2007), is often collected, used for an initial set of analyses, and then stored away on a local hard drive or network. Individually, these data sets may be small. But in the aggregate, they too are big data. The challenge is to help researchers find and use this vast pool of existing data so that they can repurpose it for new research questions and spark new insights.

Data sets that are not discoverable by anyone other than their creators are known collectively as "dark data." Dark data represents a huge cache of untapped knowledge, but finding it is challenging (Heidorn 2008).

AT A GLANCE

- Millions of data sets are underutilized because they are not findable.
- By creating communities of data sets with little human intervention, DataBridge aims to make vast troves of previously hidden dark data available for investigation and collaboration.
- DataBridge will increase the reach and enduring value of the nation's research investments by allowing more researchers across multiple disciplines to discover, share, and use data.

These data sets were collected individually, are stored in isolation from each other, and are of widely varying provenance, format, and metadata quality. In some cases, these data sets are freely available and accessible. But very rarely are they findable.

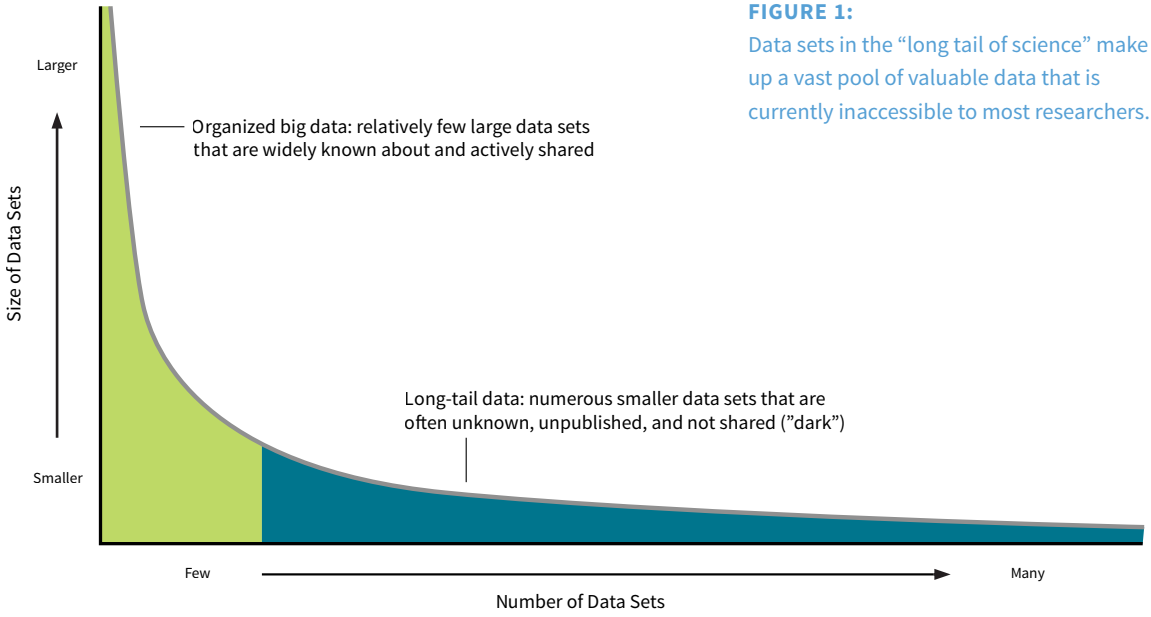


FIGURE 1:
Data sets in the "long tail of science" make up a vast pool of valuable data that is currently inaccessible to most researchers.

A platform that matches scientists with relevant, existing data sets holds tremendous potential for optimizing research investments. The median grant amount from the National Science Foundation (NSF) in 2013 was $128,700 (NSF 2013). A grant of that size is, in all likelihood, destined to produce another small data set to be added to science's long tail. Although the NSF requires that researchers submit a data management plan, those plans cover only a three-year period. Once a paper is published, that data, essentially, goes dark.

Imagine the benefits if scientists could find this dark data: Researchers' data could have a life beyond their specific project. Researchers could find other, unknown data sets that would expand their inquiries. Researchers could learn who is collecting similar data, thus fostering collaboration. Scientific research stands to benefit tremendously from a platform that builds bridges between similar data sets. In order for research funders to get the most value out of the data they sponsor, the scientific community urgently needs new tools for finding the data hidden in science's long tail.

Ideas Into Action:
Building Bridges Between Data

DataBridge aims to shine a light on dark data by making small data sets in the long tail of science more findable. Currently in a prototype stage, DataBridge is a joint project of three teams within the University of North Carolina at Chapel Hill (UNC-Chapel Hill): the Renaissance Computing Institute (RENCI), the Data Intensive Cyber Environments (DICE) Center, and the Odum Institute. These teams are part of a greater cross-institutional collaboration between UNC-Chapel Hill, North Carolina Agricultural and Technical State University (NC A&T), and Harvard University.

DataBridge is designed to allow data sets to automatically cluster together based on algorithm-detected similarities. By forming "communities" of related data sets, the system aims to enable more researchers to find and use available data sets that are relevant to their work.

There are multiple challenges to creating these communities of data. First, DataBridge must locate and assess dark data sets. Then, it must find the similarities between data sets. Next, it must cluster the data sets into communities based on those similarities. Finally, a researcher has to be able to find that data.

To solve these challenges, DataBridge is designed with four main components: the *data ingester,* the *relevance engine,* the *network engine*, and the *web-based user interface*. The *data ingester* gathers data sets' metadata from data repositories. The *relevance engine* then computes similarities between data sets. The *network engine* clusters the data sets by resulting similarities. Finally, the *web-based user interface* will enable researchers to input, search for, and view data sets based on their desired attributes. Together, these parts build the bridges needed to link researchers to relevant data.

The Data Ingester

The data ingester brings into the DataBridge system *metadata* about available data sets. This backend interface gathers metadata from thousands of scientific data sets available in repositories, individual labs, or personal computers. Currently, the ingester pulls metadata from the Dataverse Network and iRODS, two large and robust systems for data storage and sharing. The data ingester indexes multiple types of metadata, including its

scientific field, who created it, when it was created, where it was created, and what collection method was used. To improve consistency across data sets, DataBridge formats metadata into the DDI Lite schema, an international standard for describing data from the social, behavioral, and economic sciences.

The Relevance Engine

Once the metadata for a new data set is ingested into the DataBridge system, the relevance-detection algorithms compute similarities between the new data set and other data sets for which metadata was previously ingested. DataBridge has its own relevance-detection algorithms available, but also invites users to write their own algorithms. This is an important contrast to currently available search engines, which largely put scientists at the mercy of a search engine's single algorithm. With DataBridge, users can take control of their search results by writing their own algorithms based on their research parameters. The result is akin to a Google search where instead of being limited to a single search algorithm, a researcher creates her own algorithm to achieve individually-tailored results.

The Network Engine

Once these similarities have been determined, the network engine then clusters the data sets into communities via sociometric analysis. Sociometrics is a method of clustering people into communities based on their different relationships (e.g., family, friends, coworkers). DataBridge takes these principles and applies them to data sets. Each data set can be visualized as a node within an interconnected web. The connections between nodes represent various shared attributes of the linked data sets, such as content (e.g., two data sets are about climatology), method (e.g., two data sets use location-linked data from cell phones), or application (e.g., who collected the data and who else has used it). In a visualization of this network, the edges that connect the nodes vary in thickness based on the level of similarity between data sets.

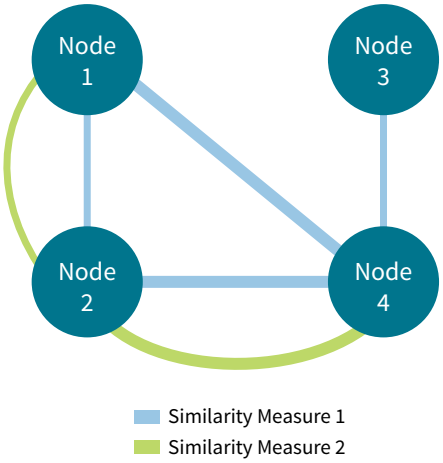


FIGURE 2:
A visualization of the connections between data sets based on different similarity measures.

The Web-Based User Interface

When DataBridge is completed, users will access it through a Web-based user interface serving three main functions. First, the interface will enable users to search for relevant data via DataBridge's existing algorithms or algorithms they write themselves. Second, users, prompted to follow the DDI Lite specifications, will be able to input their own metadata into DataBridge. Doing so will expand the reach of DataBridge and improve the relevancy of search results for the user community as a whole. Third, researchers will be able to use the interface to subscribe to automated alerts when new metadata with relevant attributes is ingested, thus making it easy for researchers to find new data sets they otherwise would never have known about.

Through its unique architecture, DataBridge enables unexpected relationships between data sets to present themselves automatically. It empowers data to make its own connections without human intervention.

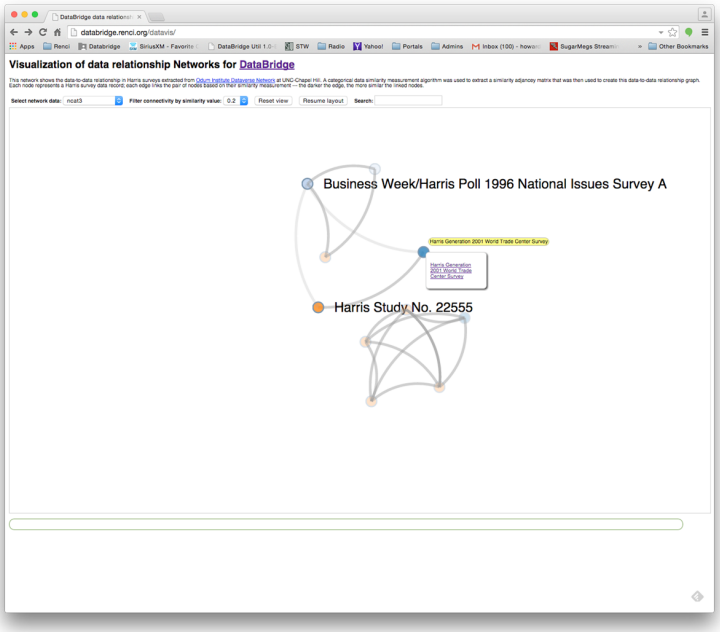


FIGURE 3:
DataBridge search results visually display clusters of data sets based on attributes that are most relevant to the user. Users can then click on any node to learn more about an individual data set or contact its owner to request access.

Because it "sees" across the entire pool of data, DataBridge offers enormous potential to connect data sets from different domains, disciplines, and countries as no other system can.

Because DataBridge is not a data repository, the system requirements are relatively modest. Each data set remains in its original repository, while its characteristics are accessed, analyzed, and connected by the DataBridge engines. This way, DataBridge can provide a streamlined system for locating data sets while each data set's home repository manages data integrity, privacy, and security.

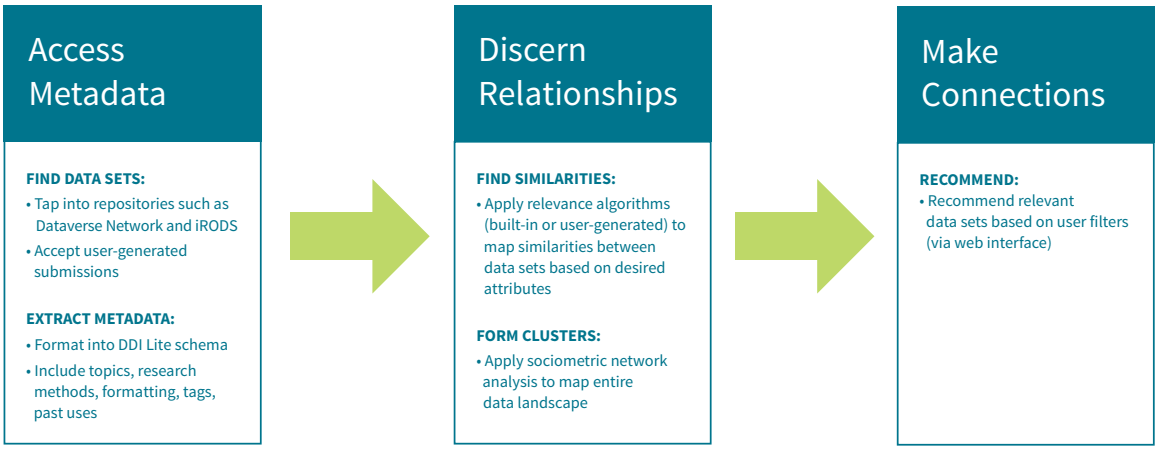


FIGURE 4: DataBridge connects data sets by pulling metadata for each data set and analyzing it to identify the relationships among all data sets in the system. A web interface helps users navigate the data landscape to identify available data sets that meet their needs.

Case Study:
Bridging SNP data sets

DataBridge is being created to meet the needs of a wide variety of researchers. One example application in genetics illustrates important aspects of the system's potential. Single Nucleotide Polymorphisms (SNPs) are genetic variations that commonly occur in a population; studying SNPs helps scientists decipher how people develop diseases or react to vaccines, drugs, and pathogens. There are several databases that collect information about different SNPs (each gene could have one or several SNPs) and their health associations, but these databases are mere repositories of information. DataBridge could allow researchers to identify similarities between SNP data sets, helping to identify data sets that can be brought together to spark new medical insights.

Using DataBridge to investigate SNPs could look like this:

1. The data ingester accesses the metadata of data sets from a variety of genetic databases to identify attributes of interest such as populations and phenotype.
2. The relevance engine finds relationships between data sets via similarity-detection algorithms that are custom built to identify relationships of interest, such as similar phenotypic characteristics or presence of specific SNPs.
3. The network engine clusters data sets into communities based on those relationships.
4. The Web-based user interface enables researchers to view the communities of data that are formed, allowing them to find collaborators working on relevant SNPs or expand their own analyses to new data sets.

The Upshot

Finding dark data in science's long tail is a complex problem requiring a complex solution. DataBridge represents promising start to overcoming a challenge that is pervasive throughout all areas of scientific inquiry.

DataBridge is wrapping up a three-year development schedule. In the first year, the project team conducted intensive research to determine the optimal architecture to make dark data findable. In the second year, the team built the prototype and ingested real data sets from quality repositories. Now in the third year, the project team is testing the system with a variety of use cases. These tests enable the team to refine the clustering results and interface to best meet the needs of DataBridge's user community.

The Big Picture

Data is often described as having a "life cycle," implying that it is born, published, and then dies. By building bridges between data sets, DataBridge extends the lifespan of data, enabling it to contribute to scientific knowledge indefinitely.

DataBridge does not archive data or collect new data. Numerous other systems have already solved those problems. The remaining problem is how to make all this data, collected in one place and stored in another, findable. Until someone can find the data, it is essentially useless. DataBridge's key innovation is making hidden data discoverable.

DataBridge gives researchers unprecedented access to dark data in the long tail of science. Researchers will gain more mileage from each data set, spark new collaborations and research questions, and enable society to reap more value from its scientific research investments.

REFERENCES

Heidorn, P. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends, 57*(2), 280-299.

NSF Funding Profile. (2013, January 1). Retrieved March 23, 2015, from https://www.nsf.gov/about/budget/fy2013/pdf/04_fy2013.pdf

Palmer, C. et al. (2007) Data curation for the long tail of science: The case of environmental sciences. *Paper presented at the Third International Digital Curation Conference*, Washington, D.C., Dec. 2007.

ACKNOWLEDGEMENT

The efforts described in this paper are funded by the NSF Office of Cyberinfrastructure OCI- 1247652, OCI-1247602, OCI-1247663 grant, "BIGDATA: Mid-Scale: ESCE: DCM: Collaborative Research: The DataBridge - A Sociometric System for Long Tail Science Data Collections," (2012-2015).

ABOUT RENCI:

RENCI, an institute of the University of North Carolina at Chapel Hill, develops and deploys advanced technologies to enable research, innovation, and economic development. For more information, see www.renci.org.

HOW TO REFERENCE THIS PAPER:

Lander, H. & Rajasekar, A. (2015): DataBridge: Creating Bridges to Find Dark Data. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0MS3QNF

renci

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

DATABRIDGE