The Shared Big Data Gateway (SBD-Gateway) project: Collaborative Archive & Data Research Environment (CADRE)

Xiaoran Yan*[1] ,Valentin Pentchev*[1], Patricia L. Mabry[1], Jamie Wittenberg[2],Robert Van Rennes[3], Matthew Hutchinson[1], Benjamin Serrette[1]
1. Indiana University Network Science Institute, 2. Indiana University Library, 3. Big Ten Academic Alliance

Project Motivation

Lack of sustainable, affordable, standardized data and text mining cyberinfrastructure for licensed, as well as open large data sets.

Solution

To address these challenges we will build a cloud based platform, CADRE, a sustainable affordable solution which leverages economies of scale through academia-industry partnership.

An essential element of our solution is establishing a community of practice. We will bring together libraries, researchers, and data providers, to address their interests and solicit their input. CADRE will facilitate sharing of data, algorithms, and visualizations; formation of standards; and research collaboration.

The Datasets

Web of Science: a leading commercial dataset, with 63M papers, 1.2B citations

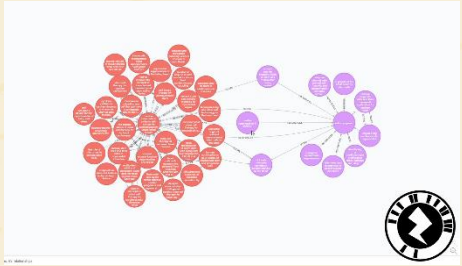
Microsoft Academic Graph: an open bibliometric dataset containing 208M documents and 1.4B citations

USPTO: 9M patent application documents

Project Goals

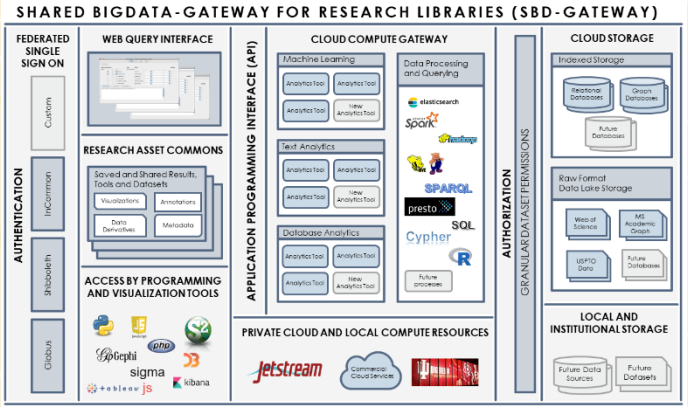
- **C**reate a community of researchers and libraries
- **A**ttract enough library partners to sustain the project beyond the 2 year IMLS grant period
- **D**eliver the initial platform at national scale with capability to host at minimum three datasets
- **R**elease as an extensible, open-source platform welcoming collaboration and future development
- **E**ngage the community to identify and prioritize additional features, datasets and other improvements for CADRE. Promote sharable and reproducible workflows, data derivatives and data standards.

CADRE Visual Query



Find out how an un-reproducible paper can impact global research. Point tablet here for an AR demo

Project Design



- **Authentication:** A federated login system utilizing institutions' proprietary authentication
- **Research Asset Commons:** A shared space in which the researchers will be able to save and if desired share and reproduce their algorithms, data subsets, derived results, tools and methods, including reproducible pipelines that can be published with identifiers
- **Compute Gateway:** A modular collection of tools, applications and technologies allowing for extensive research using cloud as well as local resources
- **Cloud Storage:** Raw Data, Relational and Graph Database storage connected to cloud and on premised compute resources

Project Partners

Contribute: We need your **user stories** for large-scale bibliometric research **http://go.iu.edu/288v**

This project was made possible in part by the Institute of Museum and Library Services LG-70-18-0202.