CollabNext - A Person-Focused Open Knowledge Graph for Collaborations with Emerging Researchers

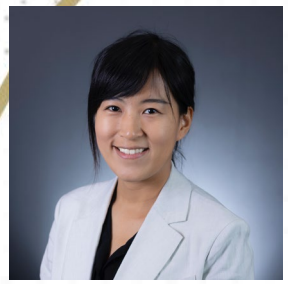
Fisk University
Georgia Tech
Morehouse College
Texas Southern University
University at Buffalo

NSF TIP Directorate
Award ID 2333737
PI: Lew Lefton
lew.lefton@gatech.edu

Georgia Tech

Leadership Team

Lew Lefton
Georgia Tech



Kexin Rong
Georgia Tech

Didier Contis
Georgia Tech

Kinnis Gosha
Morehouse College

John Porter
Morehouse College

Lila Ghemri
Texas Southern University

Craig Abbey
University at Buffalo

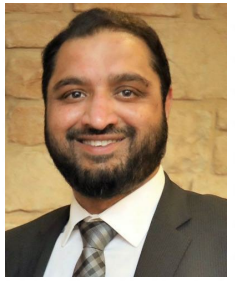
Sajid Hussain
Fisk University

Firdous Kausar
Fisk University

Lei Qian
Fisk University

A. Hannibal Hamdallahi
Fisk University

Leslie Collins
Fisk University



Sufyan Baksh
Fisk University

Georgia Tech

User Stories – Who do we care about?

- As a **principal investigator (or industry partner)**, I want to identify and contact colleagues (*at HBCUs/MSIs or at well-resourced institutions*) with interest and expertise in a specific research areas, so that I can build a stronger research team.
- As a **sponsoring agency program officer (or journal editor)**, I want to identify fresh faculty researchers who specialize in certain areas to serve as reviewers so I can expose more researchers to what successful submissions look like
- As a **program or conference organizer** I want to be curate a panel (not a man-el) of knowledgeable experts so that my event will engage a broader audience.

Georgia Tech

More User Stories – Who do we care about?

- As a **student applying to a grad program**, I want to identify researchers with whom my interests and values align, so that I can find potential advisors at an institution
- As a **graduate student or postdoc**, I want to look for potential research collaborators within and outside my institution, so that I can strengthen my research network and expertise.
- As a **program or conference organizer** I want to generate live subgraph of people, topics, and institutions based on who is in attendance at my event so that everyone can visualize new potential collaborations and existing research networks.
- As an **administrator**, I want to understand the current research capabilities and focus areas of faculty in my unit, so that I can facilitate partnerships and advocate for resources.
- As a **someone who manages a program that awards and honors researchers**, I want to identify people working in specific areas so that I can encourage them to apply and thereby creating a more diverse set of nominations and applicants

Do you have a user story which CollabNext may help support?

Georgia Tech

Project Goals – What are we trying to do?

Develop a **knowledge graph based on people, organizations, and research topics**

Adopt an intentional **human-centered design** approach which **initially prioritizes HBCUs and emerging researchers** to counterbalance the **Matthew effect**

Utilize **open science data sources** and leverage **state-of-the-art algorithms**

Georgia Tech

Project Goals – What are we trying to do?

Develop a **knowledge graph based on people, organizations, and research topics**

Adopt an intentional **human-centered design** approach which **initially prioritizes HBCUs and emerging researchers** to counterbalance the **Matthew effect**

Utilize **open science data sources** and leverage **state-of-the-art algorithms**

Georgia Tech

Invisibility in Research

Who came in 4th?

Who was the person who didn't quite make the Olympic Team?

"Invisible" researchers matter

[Katalin Karikó](#)

[David Smith](#)

Georgia Tech

The Matthew Effect – accumulated advantage

Coined by Robert K. Merton and Harriet Zuckerman
"… eminent scientists will often get more credit than a comparatively unknown researcher, even if their work is similar; it also means that credit will usually be given to researchers who are already famous."

Named from the Biblical Parable of Talents:

> For to everyone who has, will more be given, and he will have abundance; but from him who has not, even what he has will be taken away.
>
> —*Matthew 25:29*

"**The rich get richer and the poor get poorer**"
- Percy Shelley

Georgia Tech

Project Goals – What are we trying to do?

Develop a **knowledge graph based on people, organizations, and research topics**

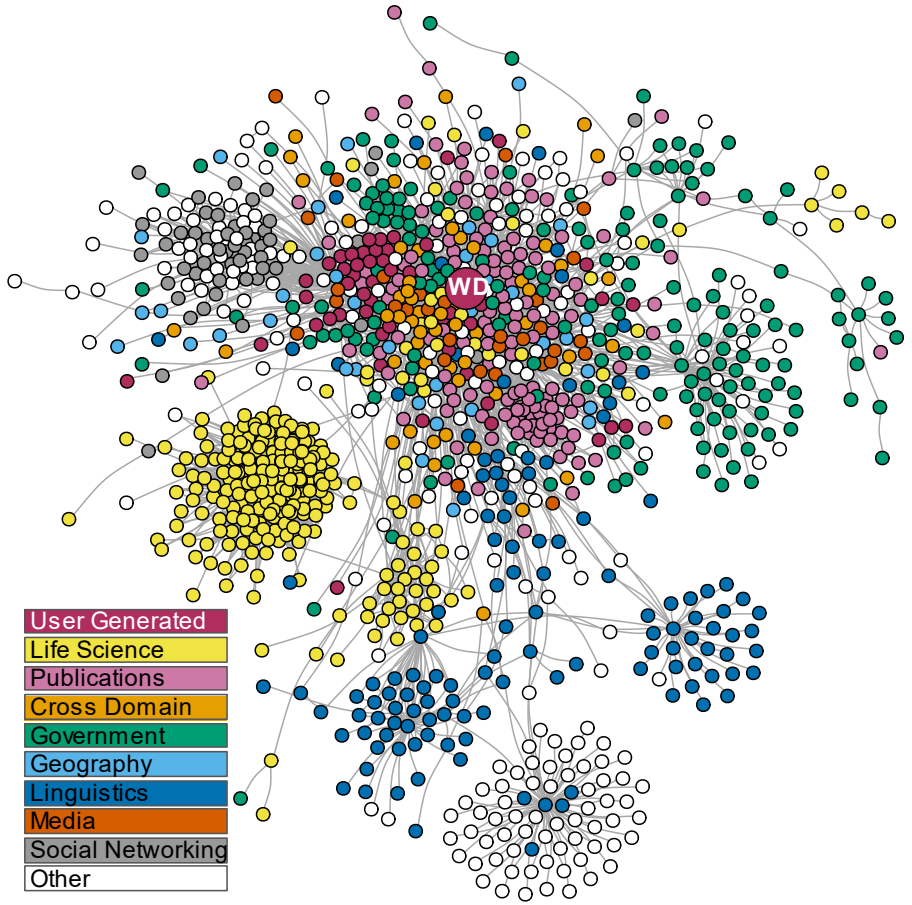
Adopt an intentional **human-centered design** approach which **initially prioritizes HBCUs and emerging researchers** to counterbalance the **Matthew effect**

Utilize **open science data sources** and leverage **state-of-the-art algorithms**

Georgia Tech

What is a Knowledge Graph?

- A **Knowledge Graph is a representation of a graph database**
- A **graph database stores linked data in the form of nodes connected with relationships**. *Compare to Relational DBs which stores data in rows and columns*
- Part of the **Semantic Web**, aka Web 3.0 (not to be confused with Web3) which is based around machine-readability and interoperability standards.
- Semantic Web is built on the idea of **Linked Data**. Examples of large linked open data sets include DBpedia and Wikidata.

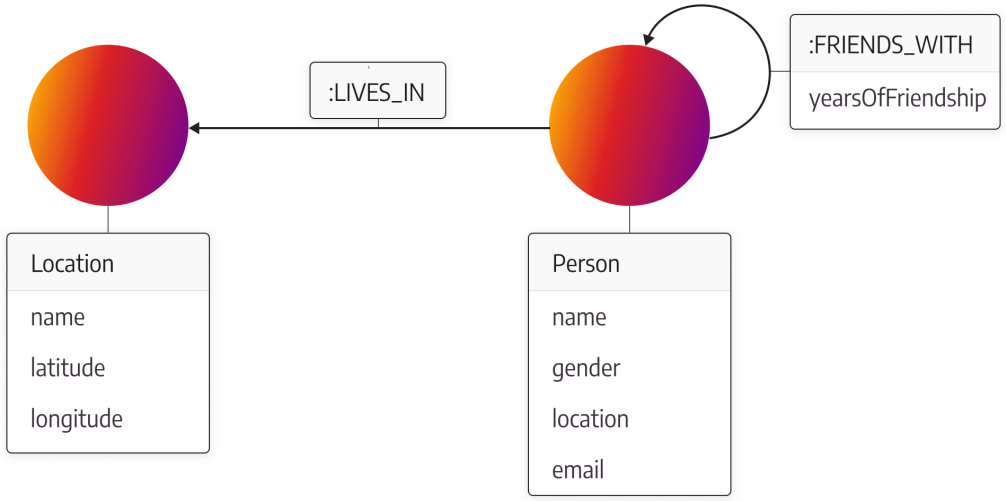


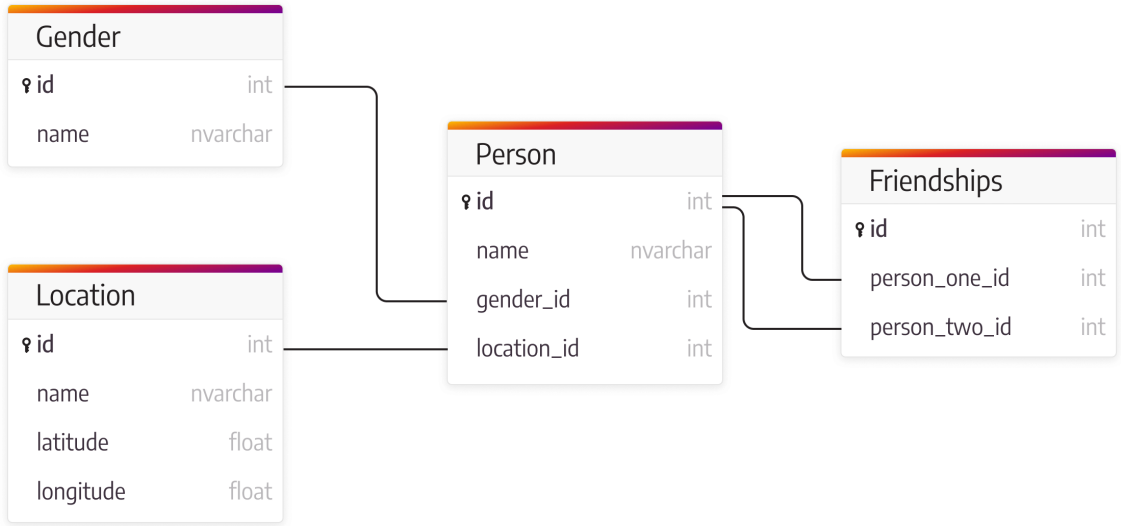
Wikidata in the Linked Open Data Cloud. Databases indicated as circles (with wikidata indicated as 'WD'), with grey lines linking databases in the network if their data is aligned.

Georgia Tech

Comparison between Graph DB and Relational DB

Compare how a Graph DB captures entities, properties of entities, relationships among entities, and properties of relationships





Query with SPARQL

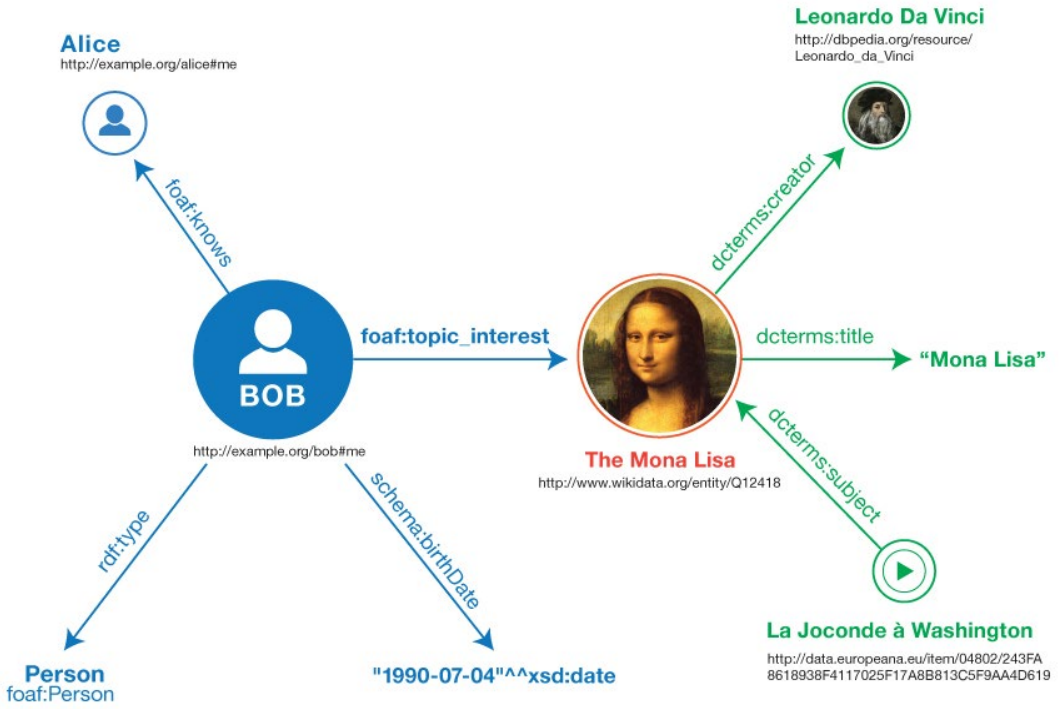
Query with SQL

Images from https://memgraph.com/blog/graph-database-vs-relational-database

Georgia Tech

Knowledge Graph Technology

- To encode the data in a KG so that it interoperable with other applications and can be linked to other data sources, the entities and relationships need to be captured in a standard structure.
- A commonly used standard is the Resource Description Framework (RDF) which captures knowledge using a subject, predicate and object called a triple.
- Other structures include JSON-LD and Labeled Property Graphs (LPGs)



neo4j ontotext MEM GRAPH ArangoDB Jena

Georgia Tech

CollabNext Example Schema

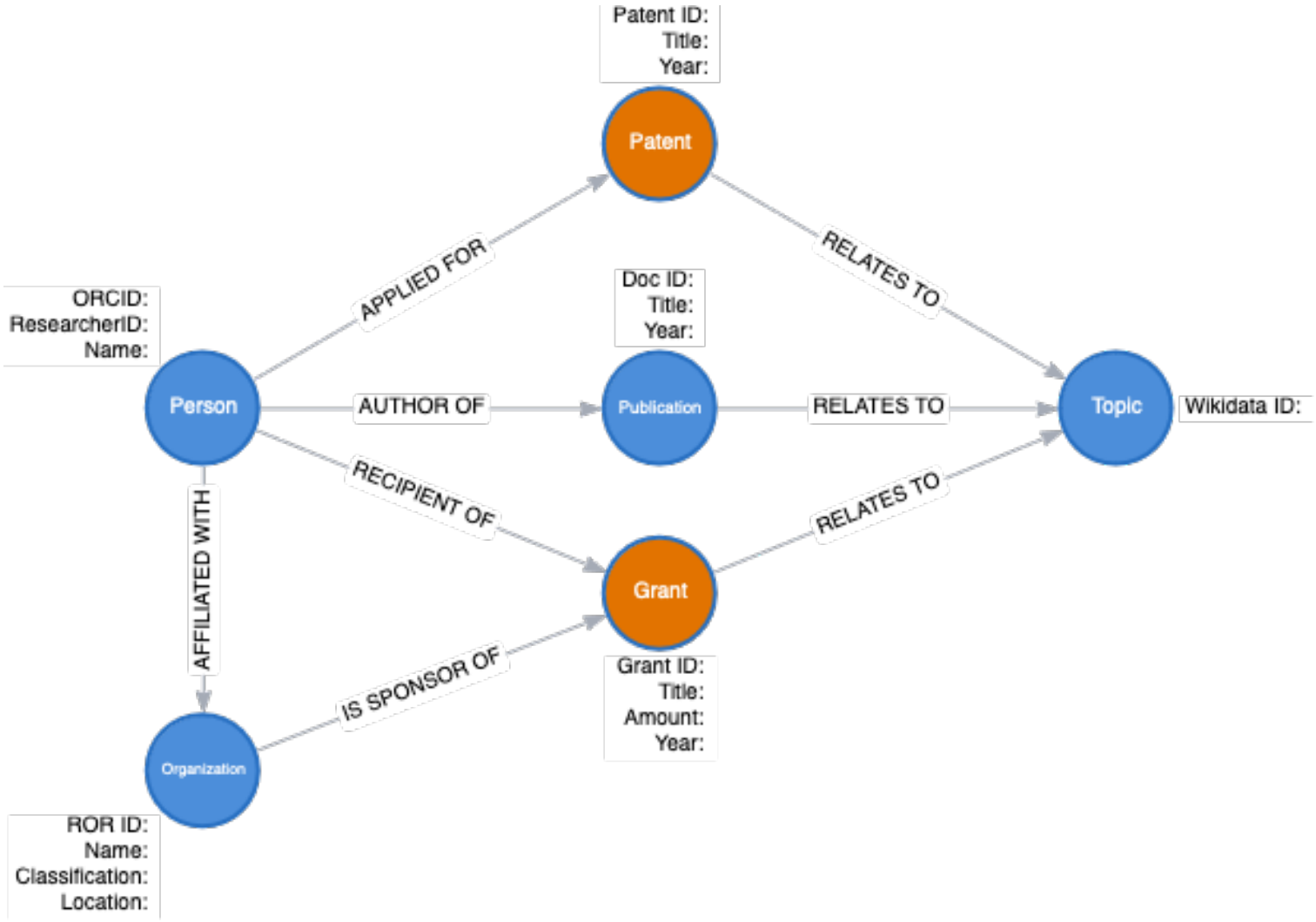
People (ORCID)

Organizations (ROR)

Topics (Wikidata)

Connecting entities

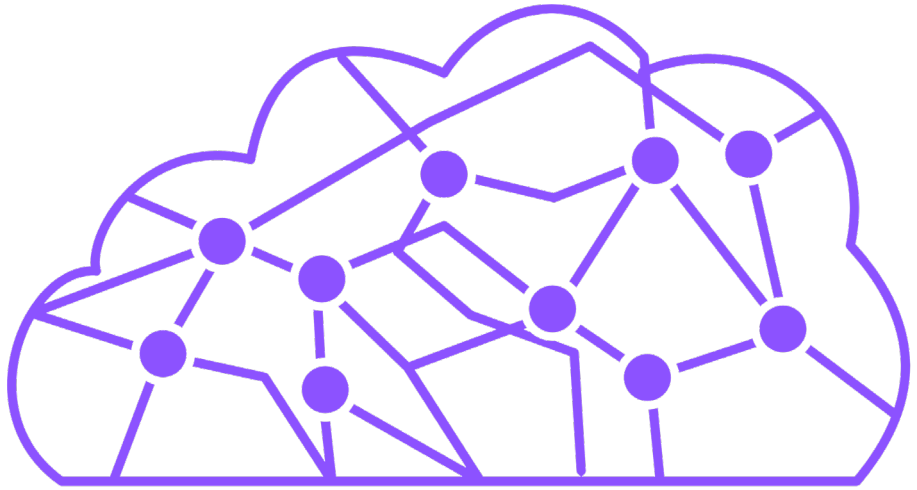
- Publications (Articles, Books, Preprints, Theses, Conference Proceedings)
- Grants/Awards
- Patents



Prototype Open Knowledge Network (Proto-OKN)

https://www.proto-okn.net/

- $26.7 million in 18 projects over 3 years invested by NSF TIP Directorate.
- This follows earlier investments through the NSF Convergence Accelerator Track A: Open Knowledge Networks
- Objectives:
  - create a publicly accessible, interconnected set of data repositories and associated knowledge graphs
  - enable data-driven, artificial intelligence-based solutions for a broad set of complex societal and economic challenges from climate change to social equity
  - empower government and non-government users — fueling evidence-based policymaking, continued strong economic growth, game-changing scientific breakthroughs,



Georgia Tech

Prototype Open Knowledge Network (Proto-OKN)

Theme 1 Use Cases (15 awards): Open data sets, both unstructured and structured, from various domains including space biology, criminal justice, environmental data, and topographic data. Most Theme 1 teams have a Federal Agency Partner.

Theme 2 Fabric (2 awards): Create an integrated data and knowledge shared infrastructure on a consistent, cloud-based "fabric"

Theme 3 Education and Public Engagement (1 award): Develop educational materials and tools aimed at various groups who will engage with the OKN

Our team from CollabNext (which is a Theme 1 project) noticed a gap between the Theme 2 OKN fabric, which provides the architecture and infrastructure, and the other Theme 1 use cases which provide the data and scientific knowledge. Namely,

Data about the people, the researchers themselves, was missing.

Georgia Tech

Project Goals – What are we trying to do?

Develop a **knowledge graph based on people, organizations, and research topics**

Adopt an intentional **human-centered design** approach which **initially prioritizes HBCUs and emerging researchers** to counterbalance the **Matthew effect**

Utilize **open science data sources** and leverage **state-of-the-art algorithms**

Georgia Tech

Open Data Sources – Persistent Identifiers

PEOPLE

Open Researcher and Contributor ID (ORCID)

ORGANIZATIONS

Research Organization Registry (ROR)

RESEARCH CONCEPTS

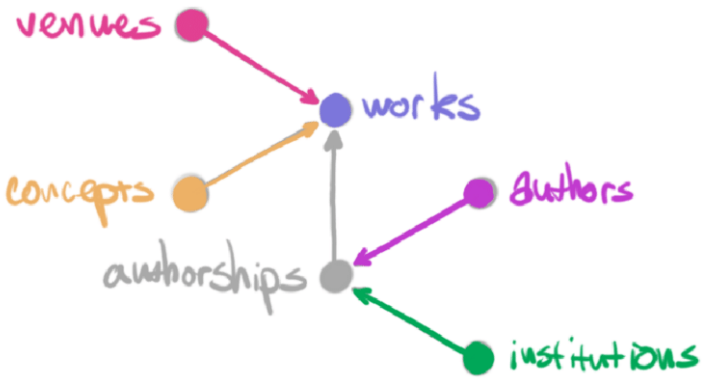
Wikidata Concepts

Georgia Tech

Open Data Sources

OpenAlex

- Formerly Microsoft Academic Graph
- Released 2022 by OurResearch
- Fully open scholarly metadata
- Data, API info, and code all released to the public
- Compared to proprietary WoS and Scopus – better coverage, less metadata (DOIs)
- More ORCID identifiers



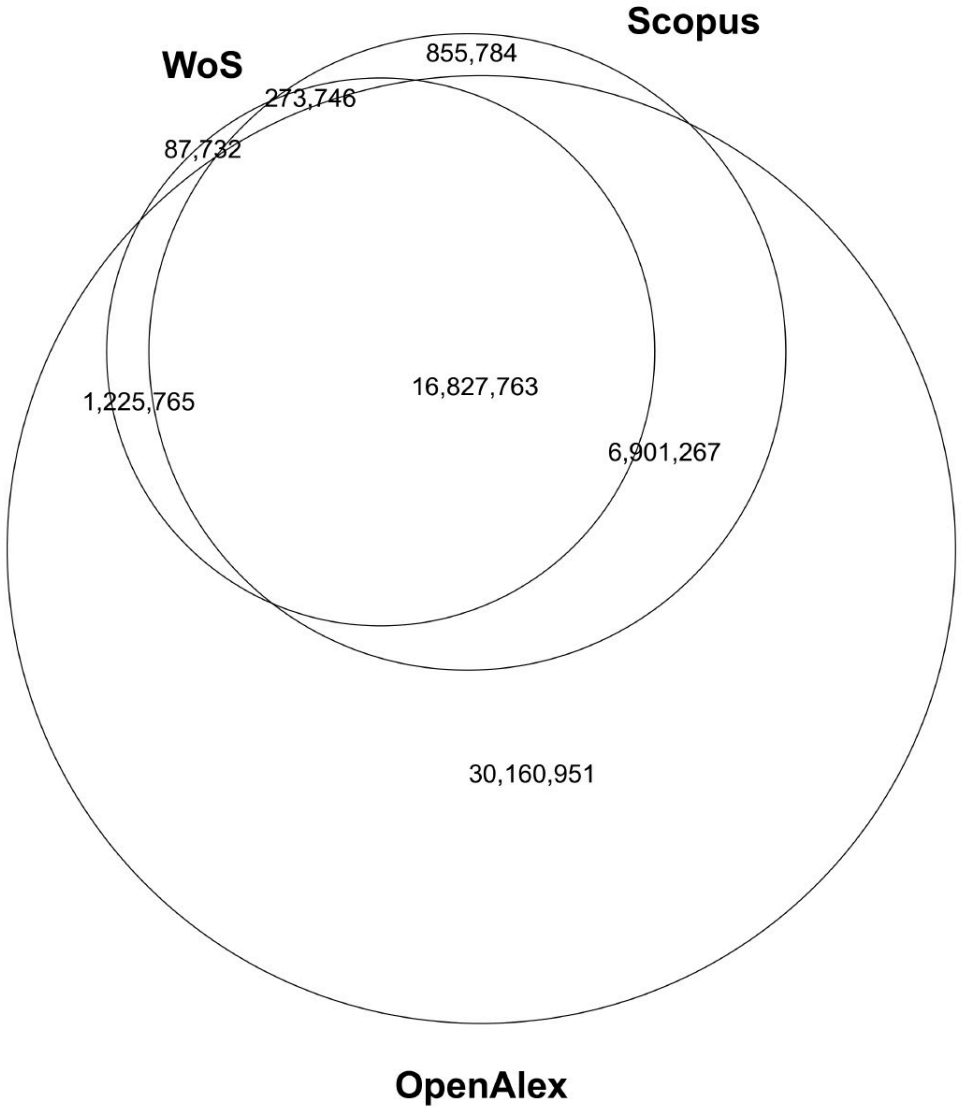


Fig. 1 Venn diagram of the intersection sizes of unique DOIs based in each database on exact DOI match, for records published between 2015 and 2022

Georgia Tech

More Data!

MUP | Center for Measuring University Performance

- Research center that collects and publishes data from several sources including NCSES (HERD, IPEDS), web scraping, etc.
- Data on Endowments, National Academy members, Faculty Awards
- Reports being expanded to include HBCUs, MSIs, etc.
- Transparent data approach (compared to US News, AAU membership, etc.)

CWTS Meaningful metrics

Centre for Science and Technology Studies (CWTS) studies scientific research and its connections to technology, innovation, and society.

- Bibliometric and scientometric tools (VOSViewer)
- Leiden Ranking

OPENSCIENCE OBSERVATORY

OpenAIRE promotes open scholarship and ensure a permanent open scholarly communication infrastructure to support European research.

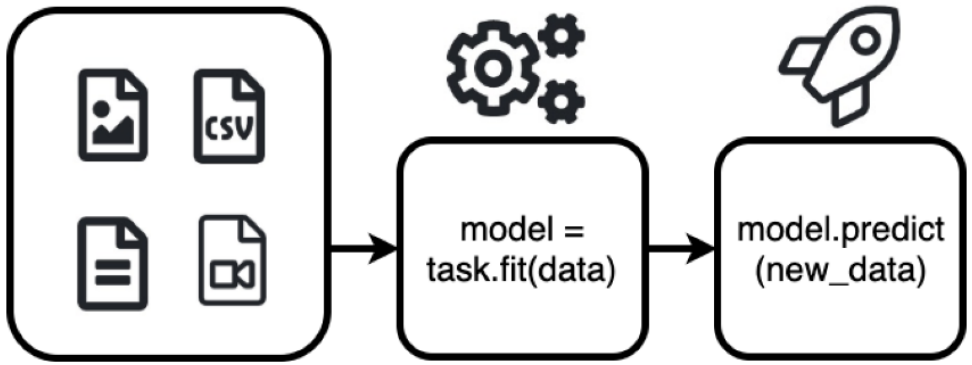
Plans to develop guidelines and data pipeline for including new and updating existing data sources:

- DVC for managing dynamic data
- Alignment with FAIR principles, CARE princples, and other standards of data ethics

arXiv DSPACE ProQuest IEEE

Georgia Tech

Advanced Algorithms



Entity resolution

- Web-Scale Academic Name Disambiguation: the WhoIsWho Benchmark, Leaderboard, and Toolkit. https://doi.org/10.48550/arXiv.2302.11848
- https://github.com/THUDM/WhoIsWho
- A knowledge graph embeddings-based approach for author name disambiguation using literals. https://openalex.org/works/w4283791844

Topic classification

- BERTopic https://maartengr.github.io/BERTopic/api/bertopic.html
- Top2Vec https://github.com/ddangelov/Top2Vec
- MFTopic https://doi.org/10.48550/arXiv.2309.01015

Other ML/AI enhancements (LLMs to write queries instead of using APIs)

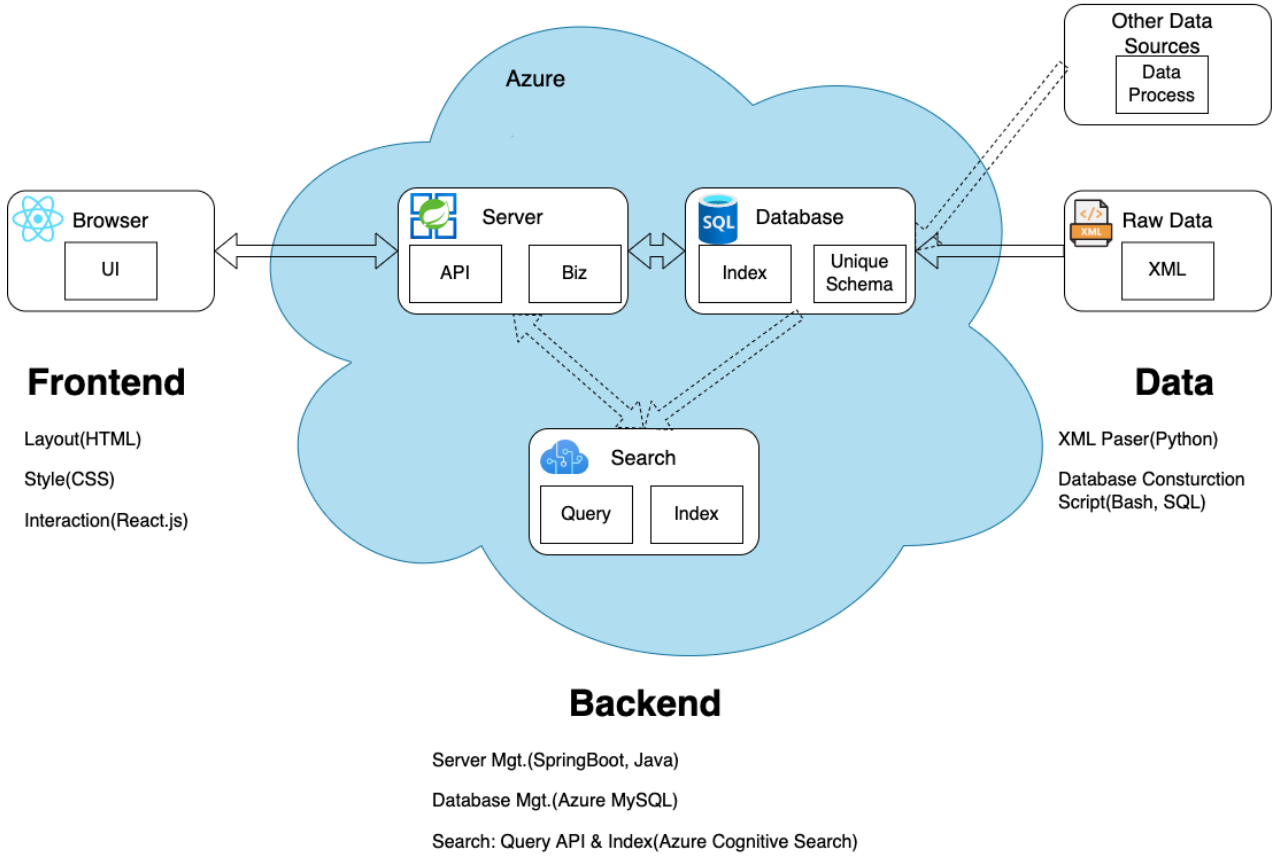
Georgia Tech

CollabNext Implementation

- Proof-of-concept running on Azure with MySQL backend.
- Includes *all* HBCUs (based on Dept. of Ed 2023 eligibility matrix)
- Modern software engineering stack (Slack, Github, Trello, Dropbox, Figma)
- Google Analytics for usage tracking (consider privacy requirements, cookie warnings, etc)

Advisory Group include representatives from:

- HBCU Faculty from Leadership Team
- Women in CS from Leadership Team
- Howard University
- AUC Data Science Initiative
- Alabama A&M
- Georgia Tech Research Institute

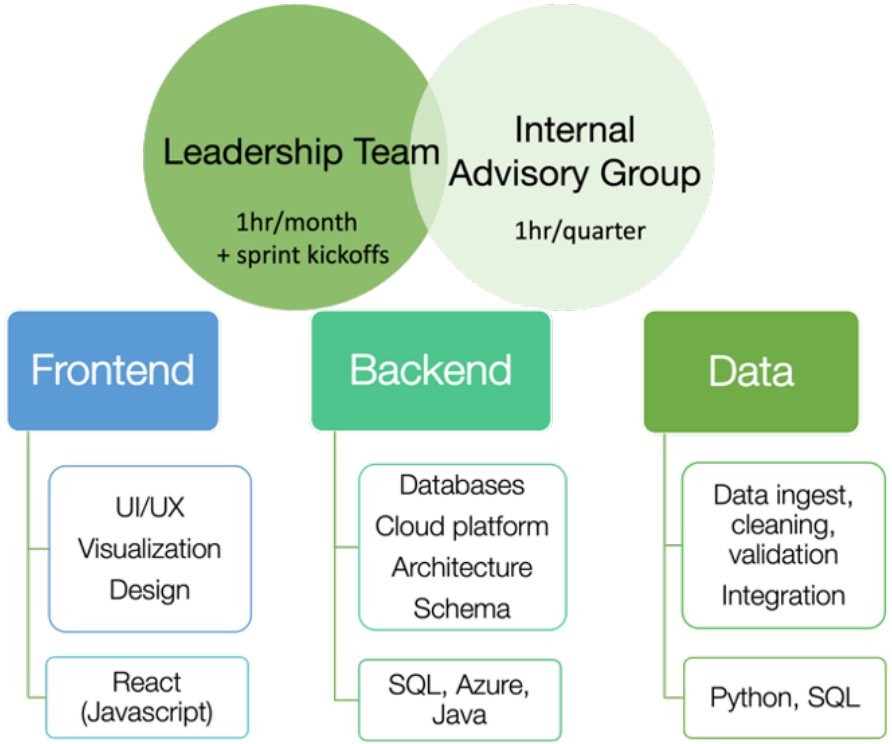


https://bit.ly/collabnext-demo

Georgia Tech

CollabNext Year 1 Goals

- Data:
  - Start integrating OpenAlex and MUP datasets
  - Finalize alpha-level schema and ontology
  - Explore Topic Classification and Entity Resolution
- Frontend:
  - Wireframes for alpha-level UI implemented
  - Basic filters and visualization
  - Advisory group input to beta-level wireframes
- Backend:
  - Infrastructure on Cloudbank
  - Relational => graph database on Theme 2 fabric
  - APIs using SPARQL
- Management and Collaboration:
  - Kickoff Partnership Team
  - Continuously seek partnerships from government agencies
  - Identify sustainability partner
  - Metrics defined,



Georgia Tech

Questions, Comments, Feedback

Demo of the Tool (if time): https://bit.ly/collabnext-demo

Contact information: lew.lefton@gatech.edu

Georgia Tech