# Biomedical Data Integration and Indexing System

Priya Deshpande[1]; Alexander Rasin[1]; Roselyne Tchoua[1]; Jacob Furst[1]; Sameer Antani[2]; Daniela Raicu[1]
[1]DePaul University, Chicago; [2]National Library of Medicine, Bethesda

## Abstract:


- Large amount of heterogeneous clinical data is generated daily.
- Clinical big data analysis is increasingly important for biomedical research, epidemiology, and education [1].
- Data integration (and indexing) systems that follow FAIR [2] (Findable, Accessible, Interoperable, and Reusable) principles are critical for enabling fine-grained access to such heterogeneous multisite data.
- We propose a design that builds on prior work in multimodal, federated, temporal data integration systems that enable indexing these data.


## Motivation:

- Re-usability of heterogeneous data from research perspective is an important goal.
- Many datasets are unavailable to other researchers because of governance, security, availability, reliability, and performance constraints.
- Need to develop methods for enabling access to such data globally, that are cost effective [3], without disturbing internal metadata structures.

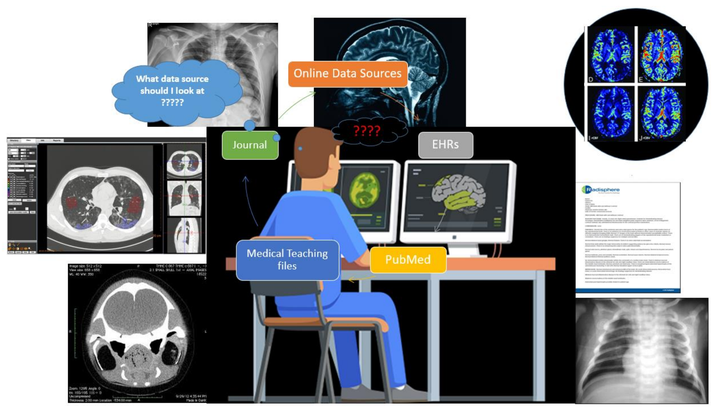


Figure 1: Motivation behind the proposed system

## Materials and Methodology:

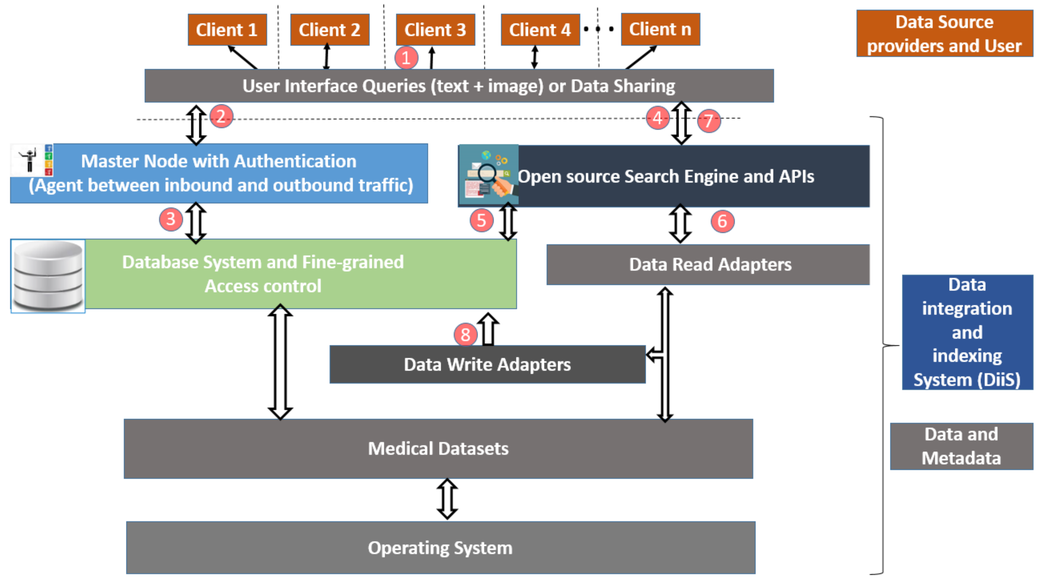


Figure 2: Architecture of Data integration and indexing System

## SWOT analysis of DiiS

- Strength: Data sharing and fine-grained access to health data.
- Weakness:
  - Not a fully automated system
  - Many stakeholders involved
- Opportunity: Collaboration across different research groups and institutions.
- Threats: Potential misconduct of donors

## Data Types in Healthcare:

- Electronic Health Records:
  - Patient's medical history, Diagnoses, Medications, Allergies
  - Treatment plans, Immunization dates
  - Radiology images
  - Laboratory and test results
- Medical Teaching files:
  - Clinical reports
  - Images
- Research Institutes and hospitals with in-house data:
  - Collection of images
  - Metadata associated with images

## Challenges in Health Data Integration:

- Heterogeneous and distributed data sources
- Data source structure and accessibility issues make indexing/search system more complex
- Process-oriented challenges: Generation, Storage, Access, and Use
- Data governance policies. Access and use controls
- Semantic and technical data source interoperability
- Performance and scalability issues

## Conclusions:

- Distributed data integration is key to advancing biomedical research.
- Massive increases in the data types, sources and velocity of data collection in healthcare industry accentuates need for data integration methods.
- Using proposed model with FAIR principles can help make data available across geographically distributed and independent organizations.

## References:


1. Hemler, Jennifer R., et al. "Practice facilitator strategies for addressing electronic health record data challenges for quality improvement: EvidenceNOW." *The Journal of the American Board of Family Medicine* 31.3 (2018): 398-409.
2. Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3 (2016).
3. Masseroli, Marco, et al. "Integrated Bio-Search: challenges and trends for the integration, search and comprehensive processing of biological information." *BMC bioinformatics* 15.1 (2014)
4. Deshpande P., Rasin A., Cao F., Yarlagadda S., Brown ET., Furst JD., Raicu DS., "Multimodal Ranked Search over Integrated Repository of Radiology Data Sources". Knowledge Discovery and Information Retrieval, Vienna, Austria, September 17-19, 2019
5. Deshpande, P., Rasin, A., Furst, J., Raicu, D., & Antani, S. (2019). Diis: A biomedical data access framework for aiding data driven research supporting fair principles. Data, 4(2), 54.


**Contact: The Visual Informatics and Data Analytics [ VIDA ] Group at DePaul**
http://facweb.cs.depaul.edu/research/vc/index.html