P20 - Data Infrastructure Approach

# Request

* **Paragraph 1:** Overall description of data ecosystem (big picture; dream big – 1 paragraph)
* Description of backend infrastructure (2 paragraphs to describe data infrastructure)
  + **Paragraph 2**: general data management
  + **Paragraph 3:**: dbt style infrastructure (long paragraph)
* Front end : 1-2 paragraphs (think salurbal
  + **Paragraph 4: Intro to approach, publications**
  + **Paragraph 5: front end**
* Integration with other products: Interactive viz products, code, papers, workshops or tutorials (SALURBAL 2 ideas)

### Project Pipeline Schemati

Graphical user interface, diagram

Description automatically generated

### Big Picture

Our data infrastructure pipeline is designed to provide machine actionable (FAIR) data/metadata throughout the entire life cycle of our research center, as shown in Figure 1. At a conceptual level, the pipeline is divided into four main stages: 1) Operationalize/Access the data/metadata 2) Understanding the data 3) Communicating our understandings and 4) integrating all of our outputs to the larger global research community through the FAIR ecosystem. Each of these stages are crucial for our research center to deliver impactful outcomes; consequently, it is vital to have a well-designed data infrastructure plan that considers the entire life cycle, or full stack, of the research project. Our approach towards implementation is that we complement traditional public research approaches with the robustness/reproducibility of software engineering best practices, scalability of industry convergent data engineering workflows, and the potential to engage audiences using modern, interactive web development frameworks.

### Access

The Access stage is the foundation of our pipeline that is responsible for intake of raw data sources and operationalizing analysis/communications ready datasets. In addition to this overall goal, there are two key directives for this stage: **integrity** and **machine actionable data/metadata**. To achieve integrity of the data we ensure original raw data sources are stored in historical storage (Drexel Servers or Azure Data Lakes\*) which are then used as inputs to the next step ETL. ETL, or Extract, Transform, and Load, involves refining raw data and reformatting it to make it more understandable, and then storing it in a central location for further analysis and use by researchers; the exact data-metadata schemas we will use will depend on data but will be informed by FAIR principles but importantly the implementation will leverage version control (Git/GitHub) to organize open source programming languages (R/Python) that integrate well with big data tools such as columnar storage (Parquet), cross-language big data structures (Apache Arrow) and if needed big data infrastructure (Apache Spark).

Now that we have operationalized our raw data as structured/machine-actionable/harmonized format we can begin transformations to operationalize datasets needed for analysis/communication; to accomplish this we utilize an industry standard data warehousing framework Data Built Tool (DBT) to organize analytics focus databases (DuckDB, Azure Synapse) to ensure our data is stored, managed, and transformed efficiently to support the research center's objectives. Briefly, DBT allows us to collaboratively modularize different stages of data transformation, starting from the raw data and progressing through several steps until the final, easy-to-understand format while providing highly accessible documentation for each step. This organized process and accessible documentation allows us to effectively handle both scale and complexity of big data transformations but also promote the transparency/accountability through the transformation process. Importantly, both our ETL and Data warehousing/DBT will be under version control for reproducibility/integrity. Taken together with historical storage, we can guarantee not just the integrity of our data but also accessibility.

Transition: Although the more data engineering methods/tools mentioned so far may be outside the scope of most traditional public health teams, they are vital to handle the increasing complexity and scale of data in public health research necessitate more advanced and specialized tools for efficient management and analysis. So one key aspect of the Data access component is to make the final data warehouse items available to researchers in a diverse set of formats (.parquet, .sas7bday, .dta) to facilitate interoperability with researcher skillsets.

### Understand

(I won’t write this as this is in the methods and not in data infrastructure).

### Communicate

We aim to not only generate scientific publications but also important to make an impact on policy. To reach the scientific community we will disseminate our research findings through both traditional scientific publications and integration into the FAIR ecosystem. Sharing our research through established platforms, such as preprints and peer-reviewed journals, ensures credibility and fosters collaboration with other researchers. Additionally, by adhering to FAIR metadata standards such as the Data Documentation Initiative (DDI) and connecting our outputs to well-established repositories like ICPSR, or WORLDFAIR’s FAIR Implementation Profiles (FIP) we contribute to a global network of shared knowledge and resources. This integration not only facilitates cross-disciplinary collaboration but also promotes the long-term sustainability and impact of our research, ultimately enhancing our ability to address pressing global challenges in public health and climate change.

To make a policy impact we need to be able reach a broader audience (non-researchers), we believe that it is essential to convey our data and findings in ways that are both accessible and engaging to non-scientific audiences. By employing contemporary technologies such as Svelte.js, D3.js, and full-stack JavaScript frameworks like Next.js, we can create interactive, dynamic, and user-friendly web interfaces that cater to a diverse audience. We also will leverage modern serverless infrastructure to (Azure Serverless Functions and Static Web Apps) to have a low cost/maintenance infrastructure to allow us to focus on delivering high value narratives/content. This web platform will also host other digital content such as workshops, videos, podcasts. By making our research outcomes more accessible and engaging, we foster greater awareness and understanding, ultimately amplifying our impact beyond the confines of academia.