P20 - data instracture v1 - rant

## Significance (why we need this)

* Having a research center is important because it enables collective knowledge sharing, **collaboration**, and access to a wider range of resources, ultimately leading to more **bigger data** and hopefully **significant research advancements**.
* **However, sucess is conditional on being able to collaborate effective particularly in the context of big data**. There are many ways this potential of a resarch center can be hamperd by data issues:
* no centralized data/metadata standards results in lack of transparency and increased friction in reusability of data; low FAIRness has the following consquences
* **Low findability** of data which limits research ideation as researchers are unable to effectively the scope of the data available.
* **Low Accessibility** of data limits pace of research as resaerchers are unable to access data in a timely fashion.
* **Low interoerability** of data provides friction for a large working group with diverse toolsets.
* Thes three ultimately translating to **Low resusability** of data within a research project which hampers the larger research effort and diverts resources away from research and into manually resolving data issues that could otherwise be autoamtated if the data/metadata was FAIR and machine actioable.
* **Importantly, these direct consequences low FAIRness scales exponentially as a project’s complexity grows with time (more people and more data) . Which is in the case of large research is 100% expected therefore it is paramount to have FAIR implmentated in to our data infsratructure and workflows**

## Innovation (what’s new about this thing we are proposing)

* Context
  + We have established above the importance of FAIR data principles. The fact is **research funders have realized this is a problem** and are now (both NSF and NIH) require DMS plans (aka FAIR implementation plans)… to ensure the proejct they fund dont fall into the above pitfalls.
  + However, the fact is that **convergent solution or implementation of FAIRness in academic research even within research domains is still a ways off**. However, there has been some progress and works to address this; these progress/current-works can be grouped into three main initiatives
    - **Profiling of how groups are implementing FAIR** across countries and research groups. This is led by WORLD-FAIR and CODATA. The first step is to collect data on the diversity of FAIR solutions and trying to find convergences and build connections… in the hope of moving the academic resarch world towards convergent solutions. (for example WP08 is a funded Fair implementation profile for SALURBAL).
    - Open-data: the predecsor to FAIR is open data movement. SO there are quite a few mature approaches towards create open data platforms (e.g. CKAN) where people can share data.
    - Metaddata-standards: However, the fact is data is largely useless without metadata… and open data is loses a lot of value if there is no global metadata standard. **One of the leading metadta stnadrards (used by ICPSR, promoted by CODATA and WORLDFAIR) is Data Documentation Initiative (DDI)**. We can adopt our machine actioable data resource to DDI standards and be confident it is FAIR within the global/time-continuum context.
  + So that leaves us all in a weird palce. Where we need to implement FAIR but there is no convergent solution… This is where our innovation comes in.
* **While academia does not have a convegent tool for implementing FAIR… industry does.**
* Big data, AI, Machine learning… are all industry buzzwords and have been for a decade or so… because industry understand they can harness FAIR data infrastructure and innoveative anaytics to make more informed business decisions to make more money. Essentially industry/companies are have resarch centers that reserach how to make more money… and because everyone wants to be efficient they developed really great FAIR implemntations that are open sourced. \*\*- So back to our quandary… we need a FAIR plan but no mature/convergent FAIR solutiosn exist in acdemia… so our innovative solution has two phases:
  + 1. we look to industry for their FIAR implmentnation to get our daata/metadata into machine actionable infrasstructure that integreates well withd ata science analytic tools.
    2. since we already machineable… it is very easy to convert to DDI to plug into the global academic FAIR ecosystem.\*\*

## Approach (what we’ll actually do)

* **Adoption of industry FAIR tools for execution as these tools are more mature and have more conevrgence than FAIR workflows in academia. This will be for within project work**
  + Machine acitoability: (R, Arrow, Parquet, DuckDB, DBT)
  + Custom interfaces: (JS)
* **FAIR principles were actually meant to be for both Software and Data. And realistically when you are on-the-ground… you cannot have FAIR data without FAIR software.**
  + Version control
  + open source - accessible software.
* **But recognition of global FAIR movement**
  + Export to globably recognized FAIR metadata standards (DDI)
  + Global FAIR repositories (ICPSR)