Project Sandiego

Red Hat OSPO

Cali Dolfi
Data Scientist

Red Hat

Problem Space: Open Source Community Analysis

- When looking at OSS projects, inside or outside the communities, you need to be able to make informed decisions
- Transitioning from metrics being a checkbox to being a tool that drives decision making

Community impact

Achieving sustainability

Risk factors can now be measured at many levels.

- Internal project health can still be measured
- Impact on community within the broader ecosystem can now be quantified
- Early detection of risk factors can inform business decisions

Business impact

Finding the ever-elusive ROI

All of these tools are well placed to help measure impact of business and community decisions.

- Targeted marketing initiatives can be tracked in a new source
- Resources can be calibrated to community health

Limitations of other tooling

- Data observability and direct access
- Visualization tooling limitations
- Most prior tooling was based on Percable and the Elasticsearch stack

Project Phases

OSS Ecosystem Mapping

- Using SourceGraph/GraphQL to identify open source projects by license
- Map connections between Open Source Projects with weighted edges (contributors, codebase etc)
- Develop plan to generate non-github projects

Explorer

- Python dash multi page dashboard
- Develop insightful metrics and visualizations to inform people about a project
- Compare other projects in the same technology ecosystem
- Build feedback loop to develop the dashboard with community input

Augur Project

- Created as the solution to data limitations
- Give git based repos and orgs, and it will load the relational database with data from the project
- Data collection tool only
- CHOASS

Phase 0

Unofficial Phase

- Discussions across RH/communities on what analysis could be useful
- Defining project scope
- Creating technical architecture
  - Dash in Python for dashboard development
- Augur data exploration with jupyter notebooks

Source:
Insert source data here
Insert source data here
Insert source data here

Current Work: Database

Augur Instance

- 10,000+ Github repos currently loaded
- Stored in the OPSO Community Cage
- Final stages of API to load repos/orgs directly

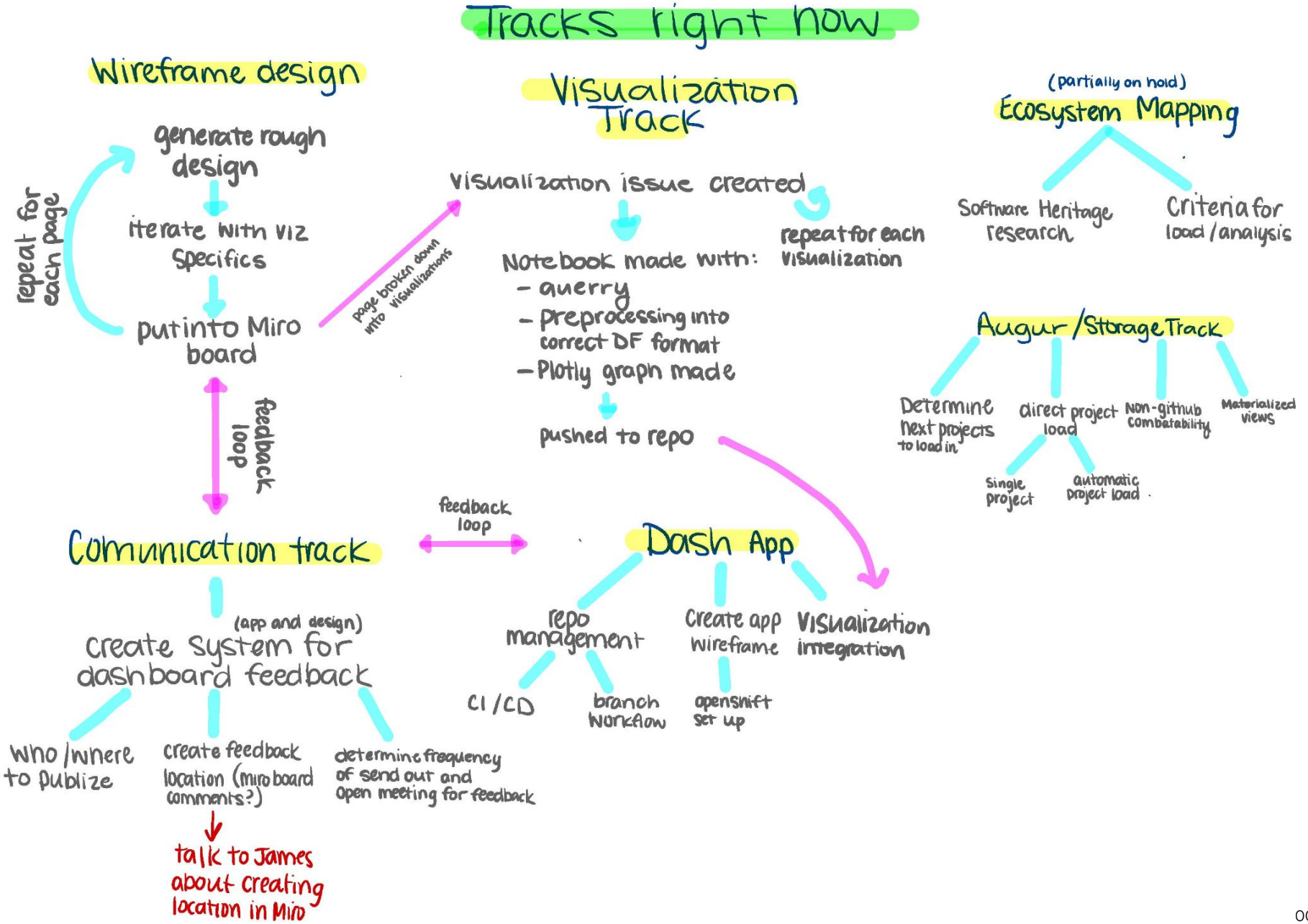
Source:
Insert source data here
Insert source data here
Insert source data here

Current Work: Explorer

Dash app

- Establishing workflow of visualization to dash integration
  - Jupyter notebooks to dash
- Technical and Design Wireframe
- Search bar

Source:
Insert source data here
Insert source data here
Insert source data here


Tracks right now
Wireframe design
generate rough design
iterate with viz specifics
put into Miro board
repeat for each page
page broken down into visualizations
Visualization Track
visualization issue created
repeat for each visualization
Notebook made with:
- querry
- preprocessing into correct DF format
- Plotly graph made
pushed to repo
(partially on hold)
Ecosystem Mapping
Software Heritage research
Criteria for load/analysis
Augur/Storage Track
Determine next projects to load in
direct project load
Single project
automatic project load
Non-github combatability
Materialized views
feedback loop
Communication track
feedback loop
Dash App
(app and design)
create system for dashboard feedback
who/where to publize
create feedback location (miro board comments?)
determine frequency of send out and open meeting for feedback
talk to James about creating location in Miro
repo management
CI/CD
branch workflow
Create app wireframe
openshift set up
Visualization integration

Red Hat

Sandiego links

Dashboard design

Github org (with project board)

High level overview timeline