Modernizing Nomad Science Data Infrastructure: An ETL Approach to Migrating Field Data

Stanley Czabanski, Pari Dar, YuFeng Lin, Ethan Peterson, Karla Vega, David Watkins

Michigan Data Consulting & UMSI-SoELL

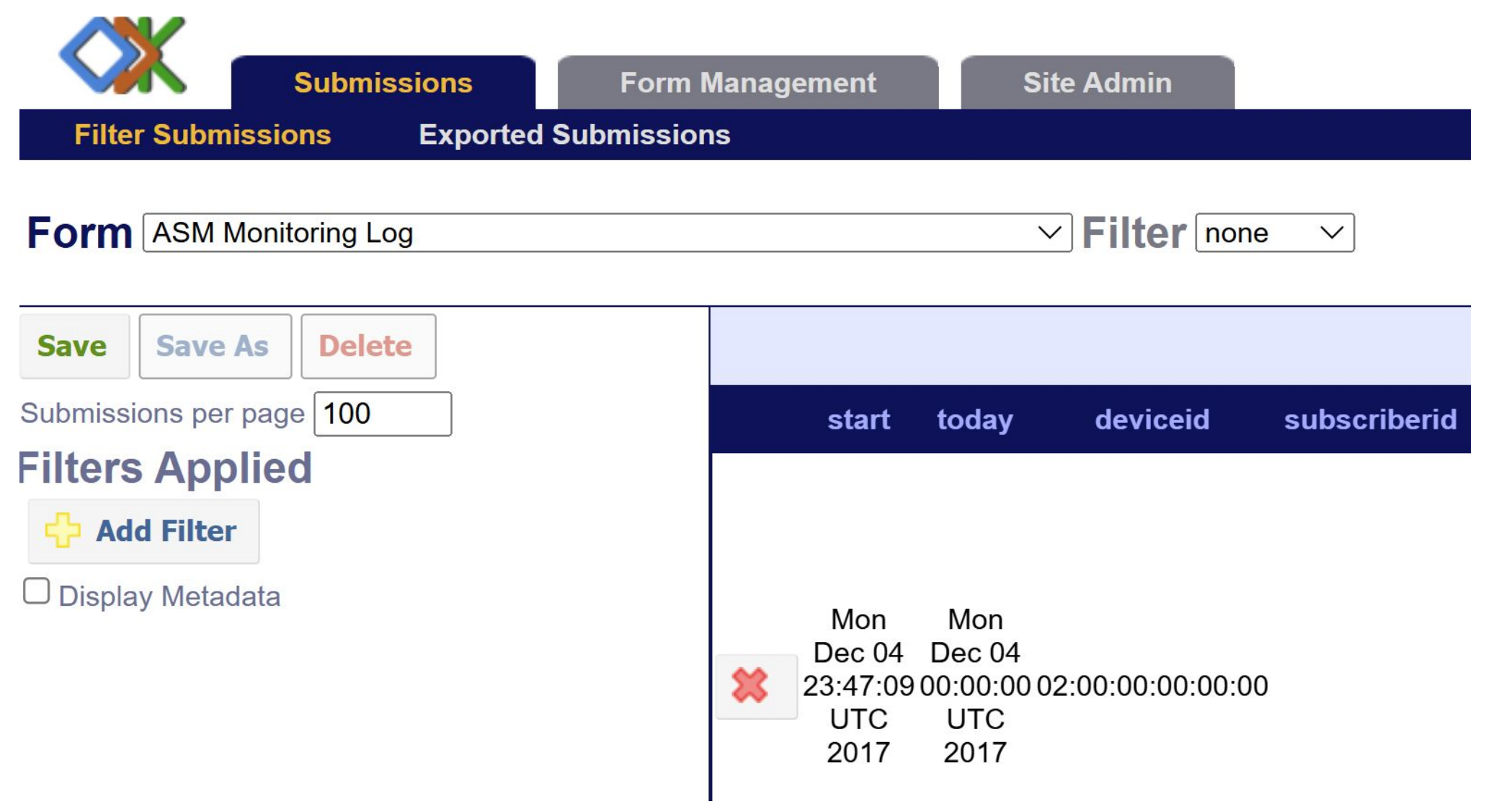
PROBLEM STATEMENT

Archaeologist team members face difficulties when working with forms due to a lack of clear documentation and tutorials, in addition to the data itself being hard to read through.

The process of exporting form data to CSV for use in Excel is inefficient, and handling images within forms poses a challenge, leading to lack of usage from team members.

CONTEXT/OBJECTIVE

As archaeologists, the members of NOMAD Science often work in remote areas which creates a reliance on ODK, a data collection server. It allows them to record data without an internet connection which can then be pushed to the database server once they hit a reconnection.



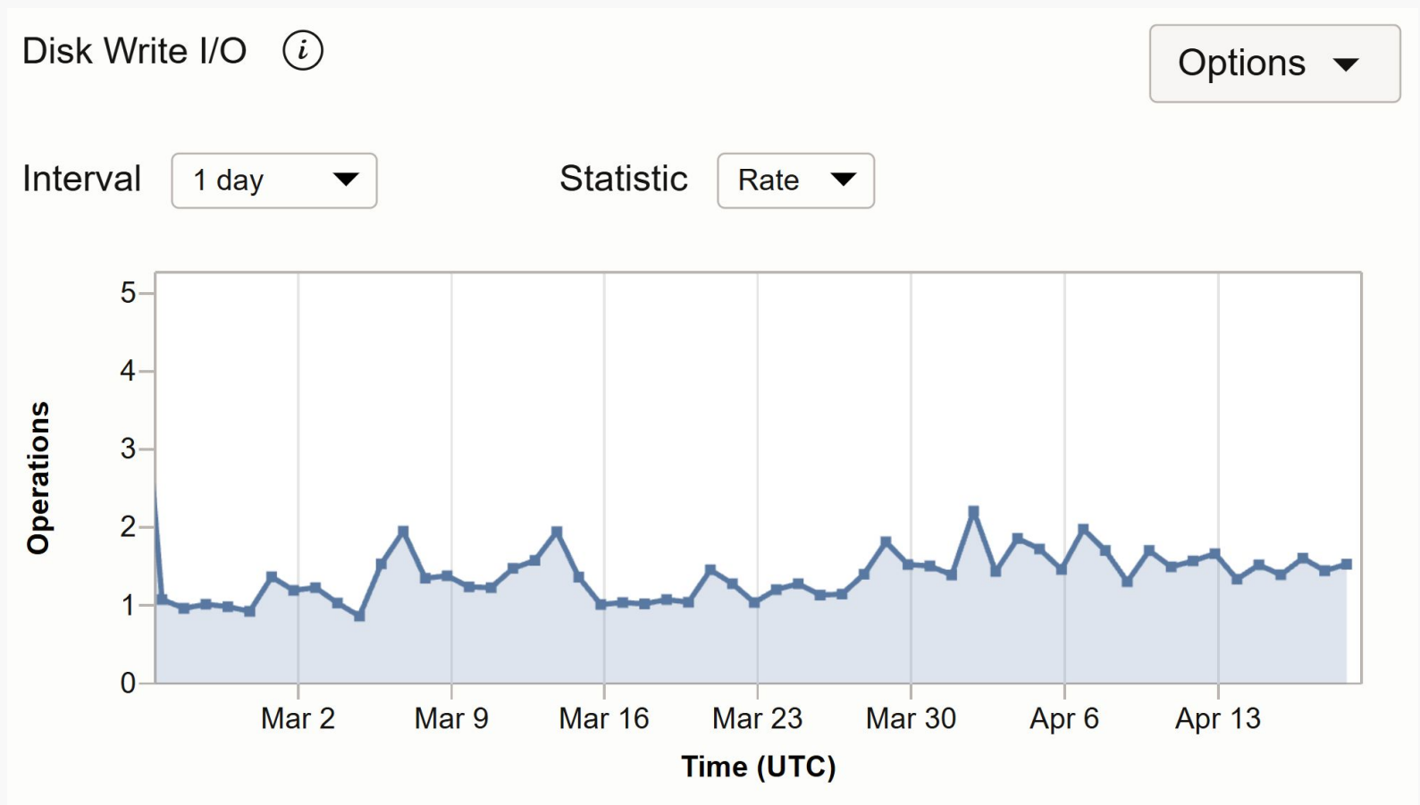
Mon Dec 04 23:47:09 UTC 2017 | Mon Dec 04 00:00:00 UTC 2017 | 02:00:00:00:00:00

When the initial manager of NOMAD Science's database on ODK moved on to new endeavors, the rest of the team was left with a lack of training to manage a large set of archaeological data. ODK houses an interface that team members have little to no familiarity with making it difficult to navigate through the dataset. This poses an issue as they cannot actually export their data or efficiently view it in order to find patterns, insights, and important characteristics. From inconsistencies in the naming conventions across different files to separate tables created from the same form, the data became too overwhelming for team members to navigate. It became too difficult to understand what most columns or values meant which created overall disparities among the proficiency levels of the team members using the dataset.

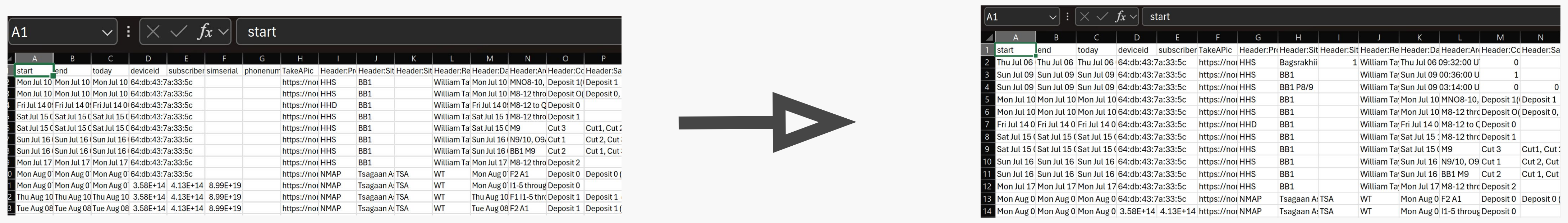
PROCESS

- In order to provide a sustainable long-term data storage solution, the team discovered through research that Oracle Cloud provides an Always Free tier for server hosting.

- Throughout the course of the project, the team has been writing data onto the Disk, with activities shown in the figure to the right, with weekly spikes as the team worked migrating the data.

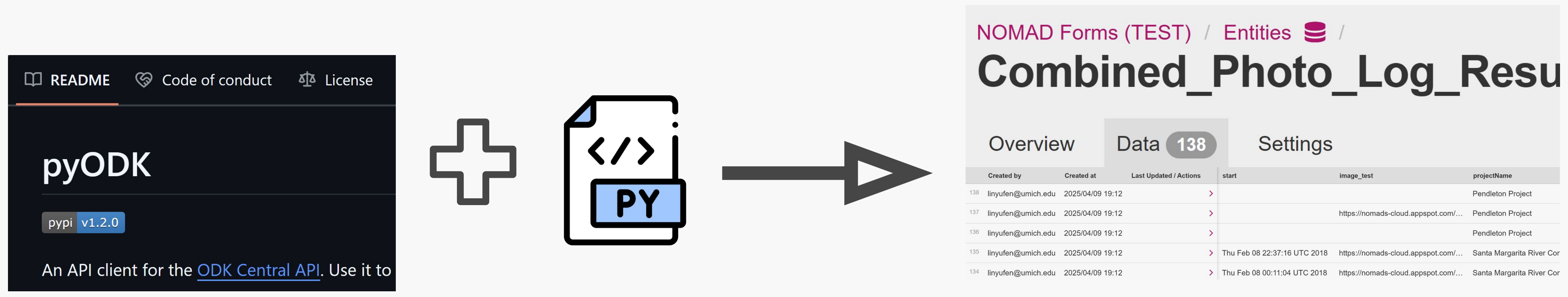


- Our task was to improve the data management system ODK Collect of for Nomad Science. As part of the ETL process (Extract, Transform and Load), our first step was to gather and review their existing files, merging duplicates and cleaning the data by removing irrelevant information, reorganizing content, and standardizing conventions.

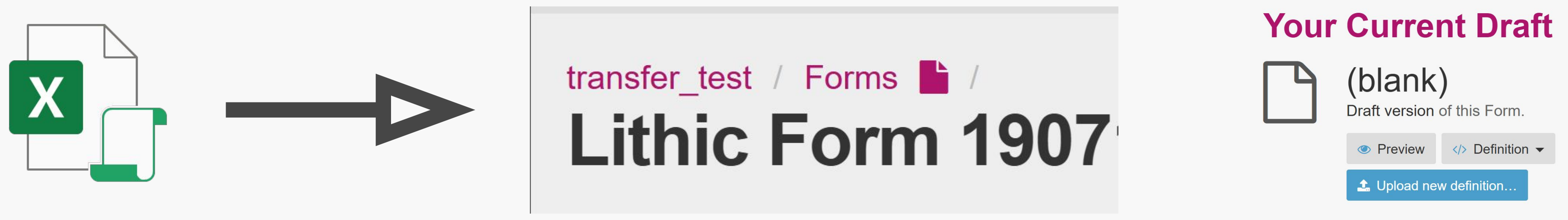


- Next, we migrated their data from an outdated server to a more modern, reliable one with a custom python script sourced from pyODK repository.

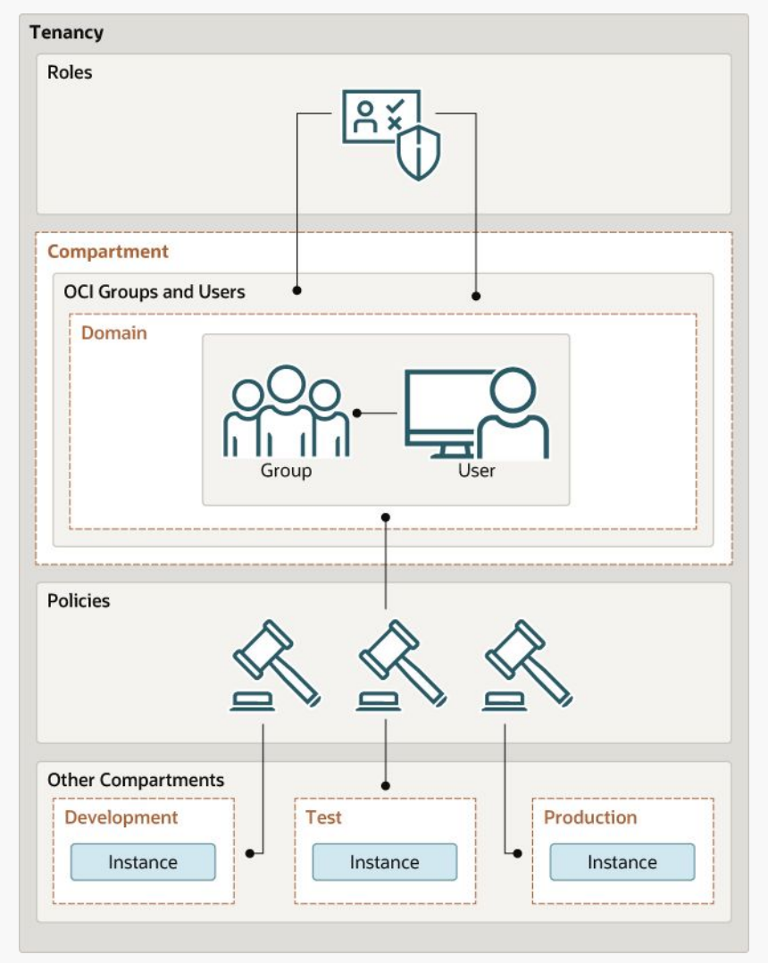
- We then moved the xlsm forms onto the new platform, so the team could better edit and distribute their forms for future use.



An API client for the ODK Central API. Use it to



- Lastly, we handed off the credentials on the cloud through user policies on Oracle Cloud. This process makes sure that our client has full access for their data without having to worry about the long-term privacy issues.



- Tutorial videos were made for the client to understand the new infrastructure.

IMPACT

By incorporating the new ODK Central platform, the team has successfully solved the major pain points from the previous infrastructure. The archaeology team is now equipped with a more user-friendly interface when collecting, editing, analyzing, and exporting data. For years to come, Nomad Science would benefit from reduced analytical time from constantly having to revisit and find their specific data. The new platform also supports PowerBI analysis with geospatial data, which are most of the entries. Mapping images with their respective metadata would require no more third party tools.

NOMAD SCIENCE

FUTURE STEPS

- Onboard members from Nomad Science using tutorials
  - Uploading from mobile devices
  - Creating a project
  - Organizing entities and entities lists
  - And more…

- Educate members on best data collection practices
  - Correct usage of data collection software so that data is more easily read
  - Present technical limitations

ACKNOWLEDGEMENTS

UM School of Information

UM College of Engineering

Nomad Science