Team Selene
Requirements Specification

Web Visualization of High Dimensionality Spatial Data

November 24th, 2016

Project Sponsor: Dr. Jay Laura, USGS Astrogeology

Faculty Mentor: Dr. Palmer

Team Lead: Daniel Ohn

Team: Zowie Haugaard, Christopher Philabaum, Kelvin Rodriguez, Makayla Shepherd

Accepted as baseline requirements for the project

Client	Team Lead

Table of Contents

1. Introduction

Planetary scientists study the planets, moons, and planetary systems of our universe, as well as their origins, formation, and processes of these bodies. Through the study of their composition, formation, and dynamics, these scientists use geophysics, atmospheric science, astrobiology, and other physical sciences to gain a better understanding of our solar system and celestial bodies, as well as our own planet Earth. Thanks to the relatively high number of observational spacecrafts currently exploring our solar system, planetary scientists have today access to a great deal of data collected from various planets and moons within our solar system, and the rate new discoveries today is very high. Yet there remain many unanswered questions within the discipline of planetary science, and in order to continue the progress of these discoveries these scientists require technologies that can assist in the analysis and exploration of often very large and complex data and problem sets.

The United States Geological Survey Astrogeology Science Center in Flagstaff, AZ was founded to assist in the training of astronauts embarking on missions to Earth's moon, and to survey and map the lunar surface. Today they are doing important research into the geology and composition of the Moon, and employ a team of geologists, planetary scientists, volcanologists, software engineers, and others to further discoveries in this area. Recently, the USGS has been given access to a data set containing hyperspectral observation data collected by the Japanese Aerospace Exploration Agency through their Kaguya lunar orbiter, also known as the SELENE lunar orbiter. This data, which in entirety is 1.4 TB in size, can through its analysis provide novel insights into the geologic makeup and composition of Earth's moon, as well as its origins.

In order to analyze this data, planetary scientists use several tools in their study. In order to receive this hyperspectral data, they must first query the database by providing the specific geographical coordinates of their predetermined region of interest. They must then download the resulting data in its entirety, and load it into a Geographical Information System (GIS) for plotting and analysis. Through this process, scientists are able to access and analyze this data and make new discoveries pertaining the geology of Earth's moon.

2. Problem Statement

To draw scientific conclusions about the formation of the Moon, as well as recent collisions, the scientists must follow a workflow that requires a lot of manual overhead. Figure 1, shown below, details the most common workflow.

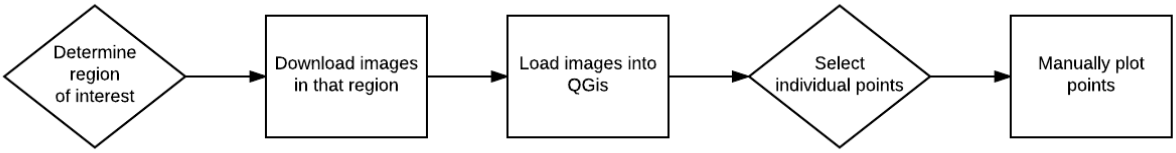

Figure 1: Basic Workflow

If a scientist wants to use the Kaguya Spectral Profiler data, they must decide what area of the Moon they want to examine, as JAXA's SELENE data archive website does not support global exploration. Unless the scientist wants to manually download the entire data set, there is way to do global exploration. Once they have decided on an area, they must then download the images associated with that area. They do this by using JAXA's Kaguya website, Figure 2, which has several search parameters. However, this website is very slow to use, as it is complicated. First, the Product Selection parameter, Figure 3, is complex, and it attempts to do product selection automatically when a product is selected. However, when a product is selected, the tool then opens more products, with no explanation. The tool does not indicate that the user must select Add or Add All to add that product to the Determination List, which is the actual list that is used in Product Selection. Another breakdown is the Observation Range parameter, which does give the user the range of data accepted, but it does not state that the Observation Range corresponds to Latitude and Longitude.

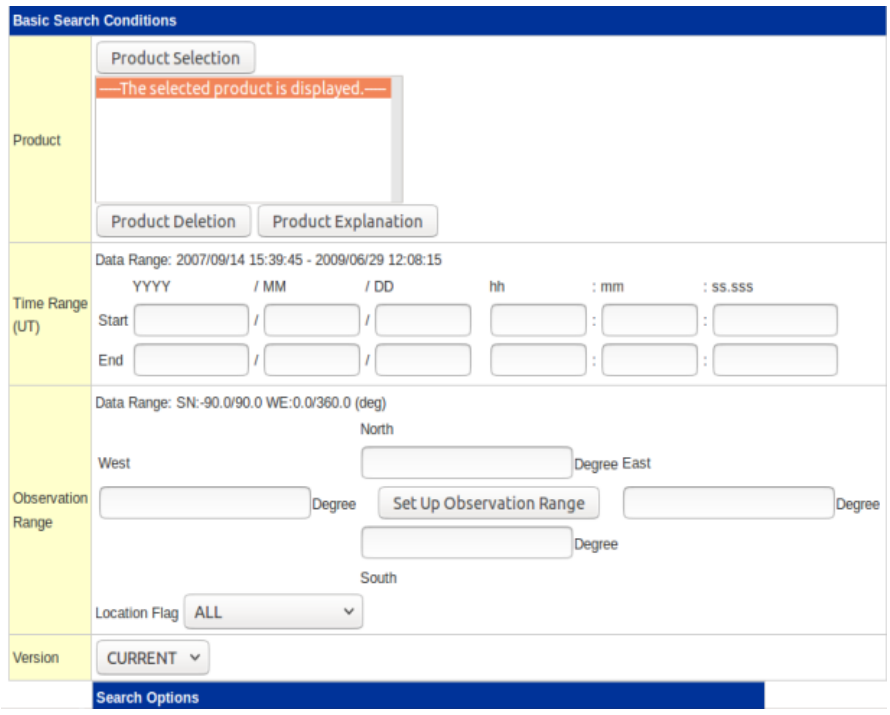

Figure 2: Website Search Parameters

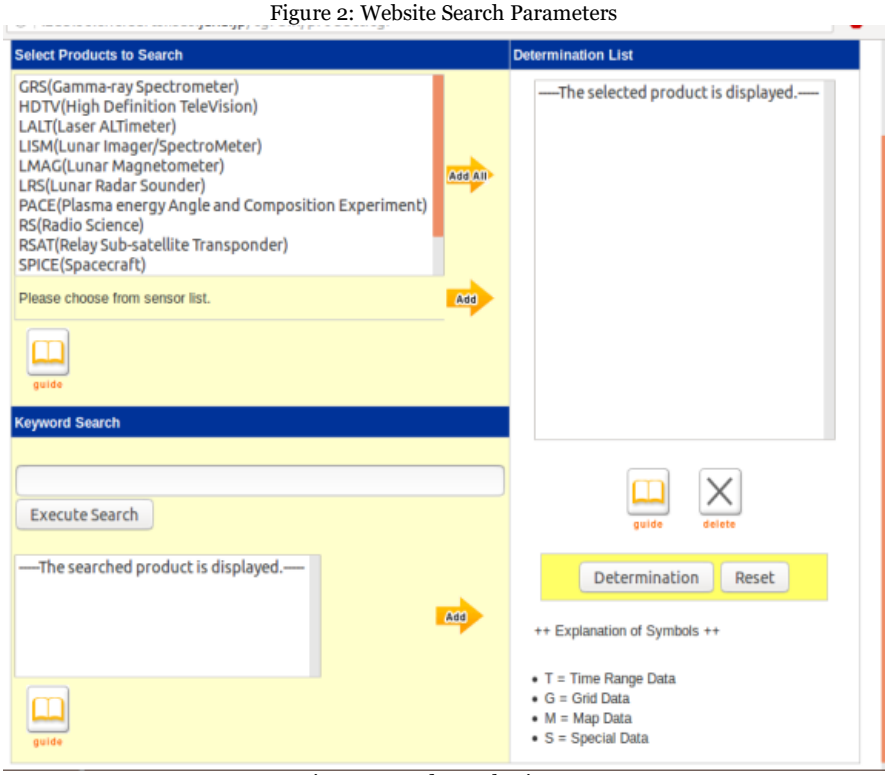

Figure 3: Product Selection

Once the user has set up their parameters, if there are more than 1000 images that fall within those parameters, the website will not show them. In fact, a scientist can only select up to a 1000 images to download, which can be a great hindrance, depending on what the scientist needs. Once the scientist has made their selection, only 10 images can be downloaded at a time using a very slow connection, making the download process very slow and time consuming.

After the images have been downloaded, the scientist will then use QGis, a planetary visualization tool, as an on-the-fly database, in order to visualize the data. Once the data is in QGis, the scientist must select an individual point, and manually plot the spectral data associated with that point. This process is time consuming and tedious, as the scientist could be following this process for hundreds, or even thousands of points. QGis, also does not allow for exploratory spatial data analysis, meaning that there is no way for the scientist to examine, or filter points based on ancillary data, such as incidence angle, emission angle, or wavelength of a point.

In summary, the current workflow is:

- Slow
- Requires manual download of only 10 images at a time
- Requires manual plotting of points
- Does not support global data exploration
- Does not allow for exploratory data analysis

Scientists using this dataset are slowed down by the challenges and breakdowns in this workflow. This workflow wastes time and money because of the large amount of manual overhead, and it can be frustrating for the scientists.

3. Solution Vision

To assist the study of this dataset, Team Selene is designing and implementing a web application for the access, visualization, exploration, and analysis of the high dimensionality spatial data captured by the Kaguya lunar orbiter. This system will greatly improve and streamline the current methodologies used in the access and analysis of this dataset, and will provide key tools to assist planetary scientists in making discoveries. Below is listed the specific key capabilities this system will implement:

- Generate a global plot of the distribution of observations
- Generate graph of observation data for each geographical point
- Allow users to explore the entire data set using actions such as pan and zoom
- Allow for the filtering of displayed data based upon incidence and emission angles, and wavelength

The specific needs of this system are explicit and as such the realization of what the solution should be was straight-forward. To allow for exploratory analysis, the user must be able to visualize the overall structure of the data set, and discover regions of interest in a more natural manner. The current methodology requires the geographical location of the data requested to be pre-defined, which does not lend itself to exploratory discovery. Additionally, this visualization needs to be viewable to the user without downloading the entirety of the data onto their own machine, meaning that a representation of the data should be easily and quickly generated and served to the user. Another drawback of the current workflow is that once the data has been downloaded, it must then be loaded into a secondary application and hyperspectral data must be manually plotted by the user. Our solution will automatically generate a visualization of the hyperspectral data when a point is selected, greatly streamlining the analysis process of specific data points.

4. Project Requirements

In order to assess the needs of scientists using this dataset, Team Selene discussed the current workflow when using this dataset with our clients. The clients pointed out some of the challenges and breakdowns with the process, and they also expressed some of their ideas for improving the workflow. Team Selene then asked the clients if there were capabilities missing from the current process that would be beneficial in analyzing this dataset.

Based on scientists' needs, as well as the needs of our clients, Team Selene, will be building a web-based application that will be:

- Responsive
- Reliable
- Interactive
- Allow data visualization and analysis
    - Globally
    - Locally

These requirements are the overall goals that the client will judge the final web application on.

4.1 Functional Requirements

While the requirements outlined above are key requirements, they are basic and user centric. Based on these key requirements, the functional guidelines below are the benchmark for success.

4.1.1 Request Map Image Data

When the application is first loaded or when parameters change a new request for map image data will be sent to retrieve the data conforming to the supplied constraints. Upon completion, the updated map data will be displayed in the main view of the application.

Request shall take the following eight parameters:

- Latitude
- Longitude
- Angle of incidence minimum
- Angle of incidence maximum
- Angle of emission minimum
- Angle of emission maximum
- Current wavelength
- Zoom level

4.1.1.1 Request Geospatial Point Data

Request shall resolve query and supply matching points relative to the specified parameters as follows:

- Latitude specifies the center of the geographic Y coordinate relative to the current request frame.
- Longitude specifies the center of the geographic X coordinate relative to the current request frame.
- Angle of incidence minimum and maximum specify the minimum and maximum angles of incidence associated with an observation that included points may contain.
- Angle of emission minimum and maximum specify the minimum and maximum angles of emission associated with an observation that included points may contain.
- Zoom level specifies the geographic boundaries of the request:
    - Boundaries are defined by relative geographic distance from the center X and center Y coordinates.
    - A smaller zoom level value defines greater distance from boundary edge to center X and Y, while larger zoom level value defines shorter distance from boundary edge to center X and Y.
    - Boundary edges define the geographic minimum and maximum coordinates included points may have.

4.1.1.2 Generate Point Distribution Plot:

A plot of the distribution of the points supplied by the request shall be generated based upon the geographic coordinates of each point. Each point shall be assigned a color dependent upon intensity of light detected at specified current wavelength. Points shall be binned according to distribution relative to image pixel density. Bins shall be colored as average of color assigned to the individual points represented by the bin.

4.1.2 Show Global Map

Function shall be called when loading or reloading the application page, as well as when individual parameters are changed. On initial application load, the parameters shall be set to the following values:

- Set latitude to 0°
- Set longitude to 0°
- Set angle of incidence minimum to 0°
- Set angle of incidence maximum to 90°
- Set angle of emission minimum to 0°
- Set angle of emission maximum to 90°
- Set current wavelength to 512.6

- Set zoom level to 0

A Request for map image data shall be sent to the server, with necessary parameters supplied based upon the current state of the client. The returned data shall be rendered by the client as a distribution map of the points matching the current criteria, and rendered as an image per the process described in section 1.1.2.

4.1.2.1 Update Image Scale

On update of the zoom parameter, the geographic X scale and geographic Y scale of the display plot will be updated as a function of the current zoom level attribute:

- A higher zoom level will update the geographic X and Y scale attributes to be a shorter geographic distance between the scale origin and axis maximum
- A lower zoom level will update the geographic X and Y scale attributes to be a longer geographic distance between the scale origin and axis maximum

4.1.3 User Control Interface

A control panel for the adjustment of individual request parameters will be displayed beside the global distribution map. This panel will include seven user input fields and two sliders for the adjustment of the individual parameters. Figure 5, below, is a mockup of the potential user interface. This interface features sliders which will be used for wavelength and zoom, and side toolbar which will contain search parameters for latitude and longitude, and incidence and emission angle.

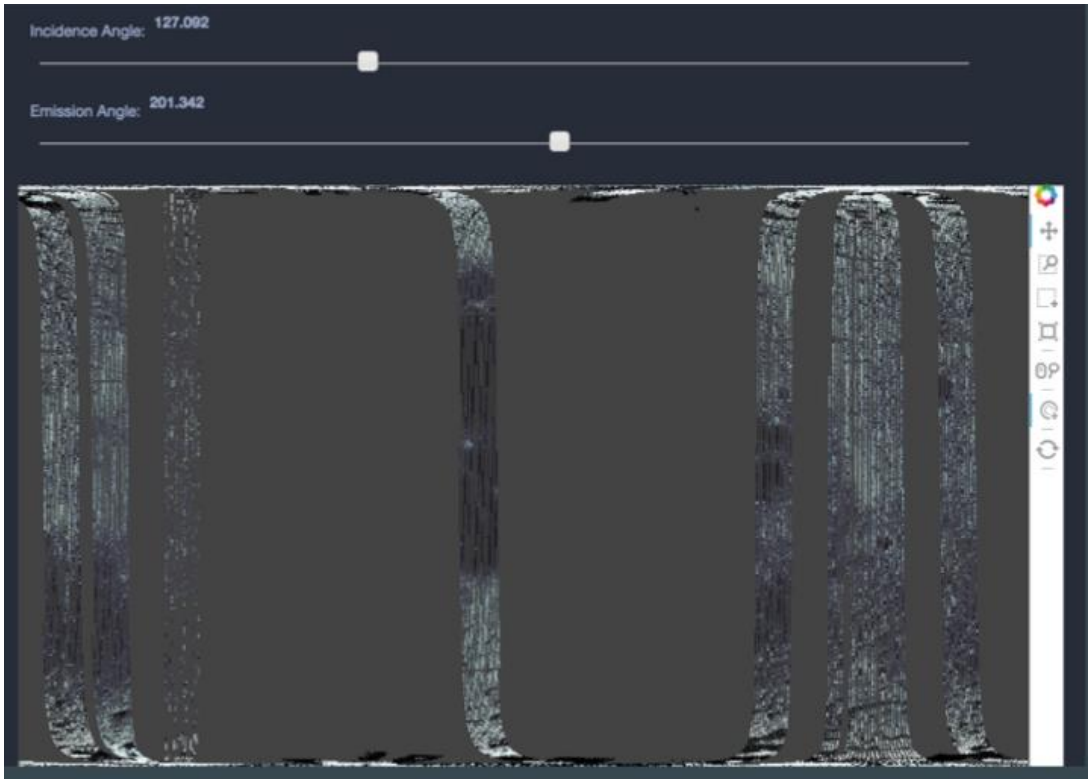

Figure 4: User Interface Mockup

4.1.3.1 Input Fields

The following parameters shall be adjustable via their corresponding input fields:

- The Latitude field shall interpret entered values as geographic Y coordinates in degrees(°). The input shall be limited to a minimum value of -90 and a maximum value of 90.
- The longitude field shall interpret entered values as geographic X coordinates in degrees(°). The input shall be limited to a minimum value of -180 and a maximum value of 180.
- The angle of incidence minimum and maximum fields shall interpret entered values as the incidence angle of observations as degrees(°). The input for both fields shall be limited to a minimum value of 0 and a maximum value of 90.
- The angle of emission minimum and maximum fields shall interpret entered values as the incidence angle of observations as degrees(°). The input for both fields shall be limited to a minimum value of 0 and a maximum value of 90.
- The current wavelength field shall interpret entered values as nanometers(nm). The input shall be limited to a minimum value of 512.6 and a maximum value of 2587.9.

4.1.3.2 Input Sliders

The following parameters shall be adjustable via their corresponding input sliders:

- Zoom level shall be an adjustable input slider oriented vertically. It shall allow for the adjustment of zoom between the minimum value of 1.0 and the maximum value of 10.0.
- Current wavelength shall be an adjustable input slider oriented horizontally. It shall allow for the adjustment of the current wavelength between the minimum value of 512.6 and the maximum value of 2587.9 The slider shall have discrete steps corresponding to each of the possible values within the collected spectral dataset.

4.1.3.3 Input Send

Upon the adjustment of any of the above listed input fields or sliders, a request for map image data shall be executed. The current value of all fields and sliders shall be used as the parameters for the request.

4.1.4 Request Hyperspectral Data

When a spatial point within the global distribution map is selected, the hyperspectral data corresponding to that point shall be retrieved. The index of the selected point shall be used as the request parameter. The vector containing the hyperspectral data retrieved at the given index shall be returned by the request. The

returned vector shall represent the hyperspectral data as key-value pairs. The key for each entry in the vector shall represent each wavelength within the observation, containing 269 discrete wavelengths observed across the near-infrared spectrum. The value corresponding to each of these wavelength keys shall represent the intensity of the light observed, ranging between 0 and 1.2.

4.1.5 Show hyperspectral data

When a spatial point within the global distribution map is selected, a request for hyperspectral data shall be sent. Upon the retrieval of the data vector, a new pane below the global map shall open. This new pane shall display a line graph plotting the retrieved hyperspectral data as follows:

- The horizontal X axis shall represent the observed wavelengths within the data set, between 512.6 and 2587.9.
- The vertical Y axis shall represent the light intensity observed, ranging from 0 to 1.2.
- A point shall be plotted within this graph for each wavelength.
- A line shall be drawn from each point to its neighbors

4.2 Performance Requirements

The performance requirements detail the quantifiable requirements that constitute success when using the application based on the key requirements and the functional requirements.

4.2.1 Guidelines for Benchmarks

Most of our requirements are very simple. The challenge is in the performance requirements considering the vast amounts of data. Our main performance goal is to maintain interactivity and maintain user attention. Any user action that takes longer than 10 seconds is considered unacceptable. For an action to be described as instantaneous, it must be finished in < .01 seconds. Additionally, because of variables outside of our programs control can hinder performance, we adopted a more statistical model for measuring performance. That is, only a percentage of user interaction instances need to stay within tolerance.

4.2.2 Requirements for Main User Interactions

- A new scene will start drawing immediately, there will be a full image rendered in less than 10 seconds 90% of the time
- Graphing the ~300 element vector in less than 10 second 90% of the time.
- Queries on ancillary data (e.g. emission angles) will be complete in less than 10 seconds 90% of the time
- The app should never lock regardless of workload.

4.3 Environmental Requirements

4.3.1 Base Requirements

The app is to be an interactive map of the moon. Users will be able to scroll, zoom and select points from the interactive application. Selecting points should present the user with a ~300 element vector plot of reflectance at that point. Zooming in at different levels should show different representations to best fit the new scale. For example, zooming into a crater should show more detailed information on the crater instead of simply presenting a resampled version of the pixel data. During these operations, the application should not freeze or otherwise terminate the user's actions. Also, if the user is panning and zooming around the map while constantly loading points onto the app, the app should maintain interactivity.

4.3.3 Motivations Behind Environmental Requirements

The goal of this project is to find better methods for plotting large geospatial points without relying on HPC solutions. That is, we want solutions that guarantee performance gains regardless of hardware setup. Therefore, solutions relying heavily on HPC are non-starters. Additionally, our clients are going to want to keep things open source. Whatever

we make should be openly available and should not depend on paid services. As our clients only care about functionality and not wide support, we have the liberty of choosing any one platform to support in trying to work out our solution. That is, we do not need to support multiple operating systems or web browsers. Likewise, choice of programming language/s is also open to our discretion. As we are building our own API, we do not need to conform to any legacy codebase.

4.4 Potential Risks

There are risks inherint with every software system. While the web application the Team Selene is building is not intended for mission critical use, we must be aware of the risks involved with the development of the system. These risks are detailed below.

4.4.1 Data Corruption

Data corruption is one of the most prominent issues for any server-client system. While the client receiving some form of corrupted would not be a mission-critical risk for anyone's safety in the foreseeable future, corruption on the data their research be based on could invalidate their claims and findings in a more extreme case. An IBM study during the 1990s showed that the likelihood of cosmic-induced error per 256 MB of RAM per month **(Odenwald & Green, 2008)**. Assuming a MB is defined to be 1000^2 bytes (as opposed to 1024^2 bytes), then there's a 1.49×10^{-15} chance of error by cosmic rays per byte per second. So by the formula for m program memory in bytes and t time in seconds, solving for t time for a 50% probability of soft error with 8 GB of memory yields a 50% of there to be a single bit error for the program within about 16 hours. However, memory cells now (November 2016) are an order magnitude smaller than those at the time of the IBM study; there's a higher concern for this type of errors with the shrinking of memory cells **(Simonite, 2008).**

By this analysis, memory errors will be a major concern the system. Like with most servers, the best way to mitigate these issues is by being sure to use ECC (error-correction code) memory modules, as well as using both a redundant array and backup system.

4.4.2 High Traffic

Another likely concern is that due to the high amount of data, it'll be relatively easier to overwhelm the server with queries. By designing the API with rate limiting, each client will only be able to make some small number of queries at any given time. This allows the server to be handle a higher volume of requests to properly manage its resources better.

4.4.3 Poor Quality Readings

With every reflectance value at a given frequency, each corresponding value has a quality score in relation to rate the quality of the reading. This means that scientists

retrieving the spectrometry at a location might be given data that could skew their results improperly. By providing a proper filter, scientists should be able to filter out such bad data.

5. Project Plan

5.1 Execution Plan

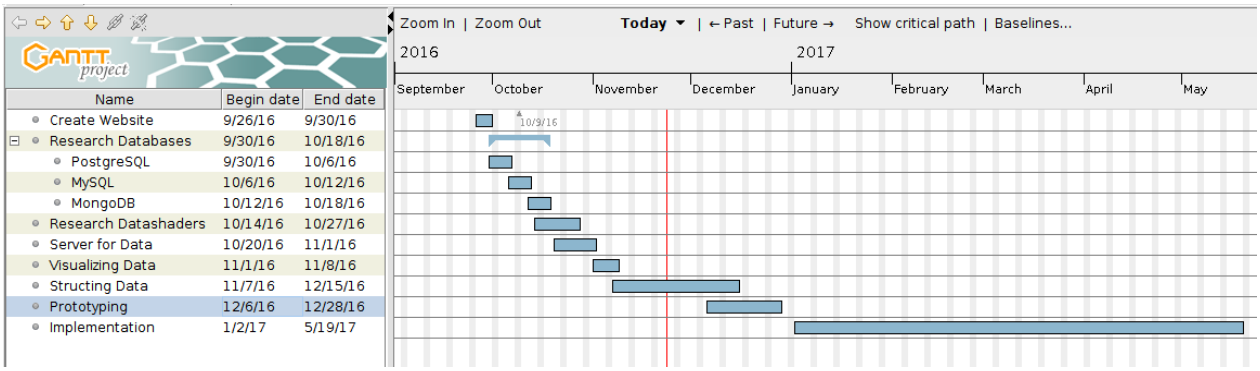

Figure 5: Gantt Chart

- Set Up Website
  - First assignment for the project which was creating a basic website that covers the team members, client, and team name.
- Research Databases
  - Finishing the first assignment, Team Selene began to research various databases to find the best method of storing their data and which one scales the best with large datasets along with querying quickly.
  - Three major databases that were researched are PostgreSQL, MySQL, and MongoDB.
- Research Datashaders
  - Once we settled on a database, we began to research datashaders and various other ways to store the images and JSON files into our database in order to send it over to the client side.
- Server for Data
  - Setting up a server and allowing all the members to be able to access it from anywhere at any time.
  - This should allow everyone in Team Selene to access the current data and be able to work on it, even if they lose the data or are unable to store it on their personal computers or laptops.
- Visualizing Data
  - With every member having access to the unstructured data, our team started the process of visualizing the data in various charts and graphs and getting a better understanding of the data we are working with.
- Structuring Data
  - Our team agreed we shall use MongoDB, but we are in the process of structuring the 40 GBs of unstructured data. This should help decrease the time it takes to query through the data.
  - Structuring the data will also allow us to easily transfer the data into a new database if we discover another database will be more beneficial for our project.
- Prototyping

    - Currently in progress, first we are creating a front-end UI for the user and implementing it with Javascript.
    - After getting a working front-end, we will set it up to the server that we setup to access the data and begin working on the backend for the web application.
- Implementation
    - The implementation process will be taking place next semester and it will take up most of the semester. We are in the structuring data and prototyping phase, but once we finish prototyping our team will have a better understanding of properly implementing correct requirements for the project.

6. Conclusion

This 1.4 terabyte data set from Kaguya's Spectral Profiler is a Big Data problem. The most common way to process this type of data for scientific purposes requires that each science center location involved in processing has its own copy of the data as there is not a better solution for the storage and visualization problem for a data set of this size. Team Selene plans to provide a web based application that delivers the spatial data in a bandwidth independent solution, and allows scientists to visualize the remotely stored data. The explicit challenges and needs of this system are non-trivial, and having a development plan to overcome them is incredibly important. As we delve into the implementation of this project the solutions and even requirements may change. Building a system that satisfies the requirements of such a product is a daunting task, and careful planning and consideration along each step of the development path is required.

Researching and finishing our previous phases of the project has helped our team get a better understand of the project requirements, along with getting a good grasp of the data that our team was presented with. Currently we are starting to prototype the software with the most straightforward approach, and then experimentation and analysis of our results. As we begin the prototyping process, issues are bound to present themselves, which we will hopefully overcome through additional research and prototyping. The key to succeeding at creating a product that suits our clients' needs will be an iterative process of refining requirements, researching solutions, and documenting our findings as best we can. If the system we develop overcomes the challenges presented in this unique problem space, our finding could indeed benefit any number of fields and projects, not just the United States Geological Survey.

Works Cited

1. Kaguya (SELENE) Data Archive. Retrieved November 20, 2016 from http://l2db.selene.darts.isas.jaxa.jp/
2. Sten F. Odenwald and James L. Green. 2008. Solar Storms: Fast Facts. (August 2008). Retrieved November 21, 2016 from https://www.scientificamerican.com/article/solar-storms-fast-facts/
3. Tom Simonite. 2008. Should every computer chip have a cosmic ray detector? (March 2008). Retrieved November 21, 2016 from https://www.newscientist.com/blog/technology/2008/03/do-we-need-cosmic-ray-alerts-for.html
4. Gantt Project. Retrieved November 22, 2016 from http://www.ganttproject.biz/