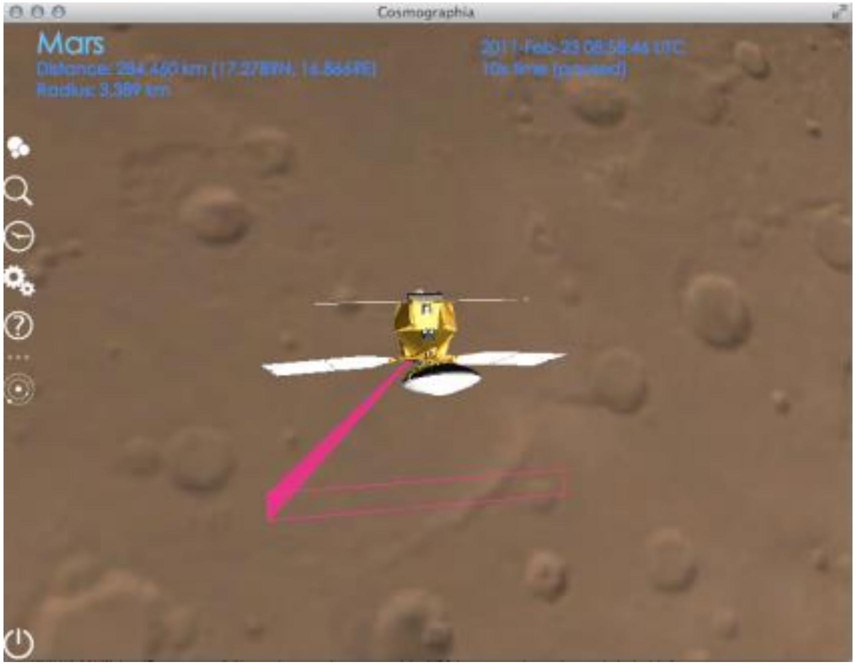
CS486C – Senior Capstone Design in Computer Science

Project Description

Project Title:	**Automatic Distribution and Synchronization Manager for Spice Database**
Sponsor Information:	Dr. Jason Laura, Research Scientist / Geographer USGS Astrogeology jlaura@usgs.gov 928.864.7366 Trent Hare, Geographer USGS Astrogeology thare@usgs.gov

Project Overview:



The NASA Navigation and Ancillary Information Facility (NAIF) group manages hundreds of gigabytes of spatio-temporal data for the position of all bodies (planets, moon, asteroids, spacecraft, etc.) in the solar system. This information is critical positional information that is used to accurately determine the position and orientation of sensors (e.g., cameras) with respect to some body. The NAIF Spice data is the foundation upon which data is accurately placed onto the surface of a body or the celestial sphere. Spice data is used by a wide variety of scientists and organizations using sensor data obtained by spacecraft. For example, the exact placement on the planet's surface of an image of Mars taken by a passing spacecraft can be obtained by computing by combining the exact time the image was taken with Spice data locating Mars and the spacecraft at that moment.

Spice data are stored in mixed plain text and binary kernels. A listing of the supported planetary missions can be found at https://naif.jpl.nasa.gov/naif/data_archived.html and a specific listing for the MESSENGER mission to Mercury is available at https://naif.jpl.nasa.gov/pub/naif/pds/data/mess-e_v_h-spice-6-v1.0/messsp_1000/. These data are traditionally stored in flat files (second link) for both download and location storage and use.

There are three factors that currently hamper access easy and efficient access to Spice data for scientists. First, the publicly available Spice data is enormous, on the order of 800 gigabytes in total. Downloading even one part of this relevant for local processing can take substantial time and bandwidth. Second, although Spice data are relatively stable, updates are frequently made to individual items by scientists using those data, e.g., small improvements and refinements. These updates are not always propagated back to the broader community. Additionally, new data are added continuously as time goes by, generating new positional data. Lastly, scientists usually need just one small element of Spice data…yet have to download entire files to locate and extract that element.

We are currently developing a tool to address the third issue, allowing scientists to specify time and body identity; the tool will locate, extract, and return exactly the positional data required. The project proposed here seeks to address the first two issues, by envisioning a distributed storage manager that works in the background to manage mirroring and smart synchronization of Spice data among any number of facilities, each of which replicates the Spice data. For example, the USGS and NAU could both offer a Spice data access API to external users. This API would allow users to utilize either service confidently because the data is being kept in sync and data integrity checks are being made.

More specifically, the goal of this project is to develop a distributed storage system that supports:

- an API for the efficient identification and location of spice kernels on the local system;
- the ability to use Spice 'meta-kernels' without modification
- an architecture that supports a single entry point (a master or name node) and multiple distributed workers (or data providers) that replicate the same data across locations (institutions);
- a method to ensure high availability between a series of distributed sites, e.g. instances of the above data and API at different locations;
- a method to ensure that data added to a single location can be 'pushed' to other locations that are mirroring;
- a means to perform data validation by polling for agreement across a majority of data providers;
- a mechanism to efficiently add new data and version existing data within the proposed solution;

Of course, our team here at USGS can provide a lot of background information and strong leads for the Capstone team to pursue in designing and implementing a solution. We expect these overall specifications to become more precise as part of the early design and requirements process.

Knowledge, skills, and expertise required for this project

- Some familiarity with the underlying architectures of distributed systems.
- Knowledge of working with heterogeneous data within a database.
- Basic knowledge of web protocols to support communication between services.

Equipment Requirements:

- We will provide access to the source data. A proof of concept implementation should be demonstrable between current laptops/desktops with modest disk space..

Software and other Deliverables:

Basic deliverables include:

- Design / Architecture documents demonstrating the system being run in as a single instance, distributed instances without a single ‘name node’, and in a fully orchestrated and replicated form.
- A working prototype that supports:
  - On disk management of the Spice kernels using the proposed solution (database, manager application on top of the file system, etc.)
  - The addition of new Spice kernels and the versioning of Spice kernels
  - A mechanism to support keeping two prototype instances in sync as new data is added or versioned
  - A mechanism to support data integrity checking between three prototypes with a ‘majority rules’ rules in place.
- Professionally documented (and tested) source code, posted to the USGS Astrogeolgoy Github page (we will make a repository for the source).