GSoC 2017 Project Proposal

Name:	Rahul Chauhan
Nickname:	rahul_world(freenode)
Country:	India
School:	National Institute of Technology Srinagar
Degree:	Bachelor of Technology
Email:	rahulnitsxr@gmail.com
Phone Number:	+91- XXX XXX XXXX
Github Profile:	https://github.com/rahulworld
Personal Website:	https://rahulworld.github.io/
LinkedIn Profile:	https://in.linkedin.com/in/rahulworld
Time-zone:	+5:30 GMT I'm comfortable adjusting my waking time to get maximum overlap in time zone with my mentor.

OSGeo Project: istSOS

Title: istSOS-Data Analysis and statistical tool suite

Introduction:

istSOS is sensor data management tool that allows collection, maintenance and publishing of monitoring observations using the Open Geospatial Consortium (OGC) Sensor Observation Service (SOS)standard.

OAT (Observation Analysis Tools)can be used as a preprocessor for calibration observations, integrating the creation of observations for calibration directly from sensor time-series. The tool consists in an expandable Python library of processing methods and an interface integrated in the QGIS FREEWAT plug-in which includes a large number of modelling capabilities, data management tools and calibration capacity.

The OAT library is an expandable Python library that can easily interact with istSOS.

The goal of my project primarily is to add OAT extension in RESTFull Web api and OAT extension having data analysis and statistical tools for istSOS which will be used to automate the creation of statisticat documents using OAT library and harvesting the data from an istSOS server and then publishing the result on Open Data Portals(CKAN).

Mockups:
Adding OAT extension view.
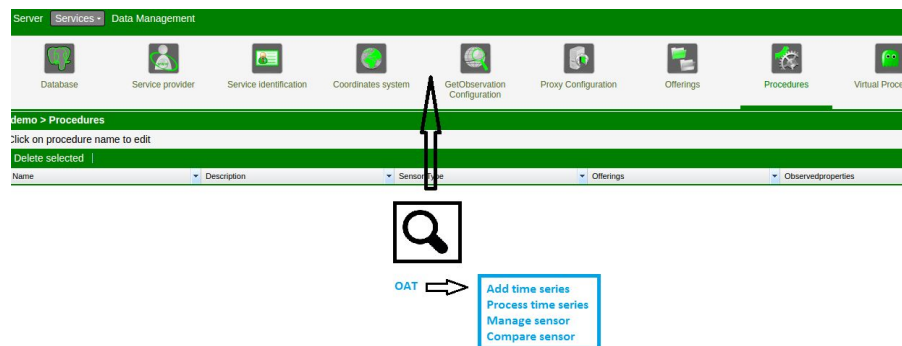
final view of the publishing the result on ckan.
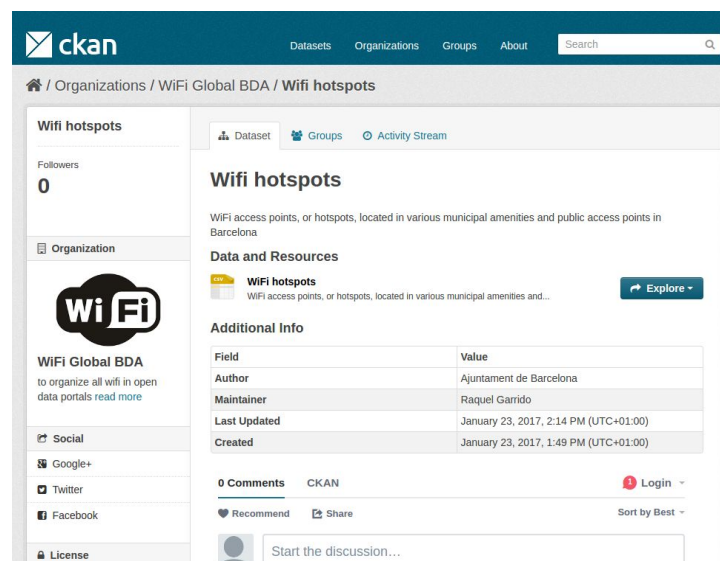
Idea:

istSOS has gone through a lot of improvements since 2009. This has somehow affected the simplified interaction with measurements and SOS service settings.

The aim of my project primarily is to create a data analysis and statistical tools suite which will be used execute time-series visualization, pre-processing of data for model development and post-processing of sensor observations data. The data analysis and statistical tools suite will cover all the features of OAT library.

- **OAT(Observation Analysis Tool)**
  OAT is designed to facilitate the import, analysis and visualization of time-series data and the use of these data to support model construction and advanced model calibration.

- **The OAT library implements two class:**

1. Sensor class that is designated to handle time series data and metadata.
2. Method class which is designated to represent a processing method.
   OAT can retrieve data stored in local files or database(CSV, istSOS, MODFLOW's hob file, MODFLOW's gage file etc).

- **currently available OAT.method:**
  - **analysis method on raw data:**
    - **Resampling:**
      - OAT takes advantage of the PANDAS (McKinney, 2010), NUMPY and SCIPY (Van der Walt et. al. 2011) packages.
      - Pandas use ***upsample/interpolation*** time series data to a higher frequency and interpolate the new observations.
      - Pandas use ***downsample/Aggregation*** time series data to a lower frequency and summarize the higher frequency observations.
      - In the case of upsampling, care may be needed in determining how the fine-grained observations are calculated using ***interpolation***.
      - In the case of downsampling, care may be needed in selecting the ***summary statistics*** used to calculate the new ***aggregated values***.

    - **Regularization**:
      - Regularization refers to the method of preventing overfitting, by explicitly controlling the model complexity.
      - It leads to smoothing of the regression line and thus prevents overfitting. It does so by penalizing the bent of the regression line that tries to closely match the noisy data points.

- **Data interpolation:**

  I have used data interpolation in my last project using B-spline interpolation. Interpolation is typically used as preprocessing step. Interpolation can be done using many method (spline, kernel smoothing method, wavelate transformation).

- **Fitting:**

  fitting is matching 2 results (ie measuring error rate of datasets, points, n-order curves). the first result is the output of the observed system or a desired behavior. the second one is learning system's output.

- **Validation:**

  By splitting the data in multiple parts we can check if an analysis (like a fitted model) based on one part of the data generalizes to another part of the data as well.

- **Filling:**

  Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data.

- **Data quality assessment , Decomposition or filtering on time-series(low, medium, high frequencies), Hydrograph separation, Evaporation etc.**

  OAT.Method objects are based on TSPROC processing capabilities. TSPROC(Time Series PROCessor) is a software package designed to assist in the calibration of models by editing and distilling time series datasets into more meaningful observations to be used in the optimization objective function. TSPROC uses a simple scripting language to process and analyze time series. TSPROC has perform calculations on time-series data commonly associated with surface-water models, including calculation of flow volumes, transformation by means of basic arithmetic operations, and generation of seasonal and **annual/period statistics** and hydrologic indices.

  - **Statistical Tools**:
    - **Aggregation:**

Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis.

- **Exceedance-time:**
  Cumulative frequency analysis is the analysis of the frequency of occurrence of values of a phenomenon less than a reference value.

- **Summary and Period statistics:**
  - *Summary statistics* are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible.
  - Measure of location, or *central tendency*, such as the arithmetic mean, Measure of statistical dispersion like the *standard deviation*, measure of the shape of the distribution like *skewness or kurtosis*.
  - If more than one variable is measured, a measure of statistical dependence such as a *correlation coefficient*.
  - *Period statistics* is time series statistics on a monthly or annual basis without the need to supply a separate dates list.

- **CKAN:**
  - CKAN software is the most popular data portal software and also provides a framework for writing custom harvesters to import data in any formate from server (istSOS server) for publishing. it is highly customizable, allowing to define default tags, groups, users and permissions for the imported datasets. Once data is published, users can use its faceted search features to browse and find the data they need, and preview it using maps, graphs and tables.

- **Exposing data and automate metadata:**
  - For each sensor, the "metadata section" includes name, description, location (lat, lon, elev), unit of measure, observed property, coordinate system, time-zone, frequency, weight statistic and data availability (time interval).
  - Meta Manager can be directly harvested by CKAN software to streamline populating an organization's data directly into a CKAN-based portal.
  - Metadata records that reside in a relational database (**Postgres**).
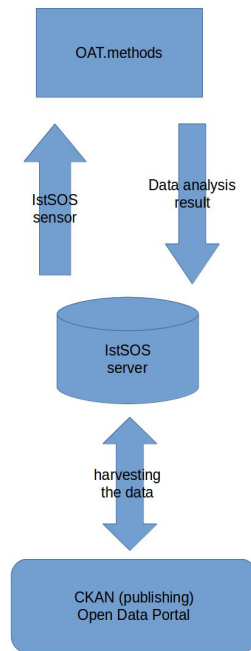
Data analysis and publishing data on CKAN

How I intend to continue being an active member of your project and/or istSOS AFTER the summer is over:

After the successful completion of the project, I will continue working for istSOS and developing new modules in this area. In future work, I have also suggested some topics that can be implemented and increase the functionalities of istSOS and create pull request for istSOS android-demo [2] currently working on istSOS java-core, android-demo and want to develop new module.

Timeline:

Before May 4: Self study

- Familiarize myself completely with istSOS functions and architecture.
- Also discuss with mentor, a plan, to do performance improvements.

May 4 – May 15: istSOS Community Bonding Period

- To do self coding with istSOS to improve my further understanding and ease of use with the software.
- Experimentation and testing with istSOS plugin development.

- Testing them on different dataset.

- During this period, I will remain active on IRC and Mailing List to discuss and finalize on how to proceed with the project.

- Create a github repo where all the code will be pushed once the testing of implemented algorithms is done.(https://github.com/rahulworld/).

May 15 – May 30: istSOS Community Bonding Period
- Start coding early from 2nd week of May.
- Design GUI for all activity(extension and sub extension).

June 1- June 7:

- Adding in GUI(OAT extension and its sub-extension).

- Implementation of OAT extension using OAT library.

June 8- June 14:
- Testing OAT library with istSOS sensor.
- debugging.

June 15 -June 22:

- Implement OAT.method i.e. resampling,regularization, data interpolation, fitting, filling.

June 23 -June 30: (Phase 1 Evaluation)
- GUI to have OAT.method functionalities i.e. resampling, regularization, data interpolation, fitting, filling.

- Test the working of these operations.

- Testing the overall working of each and every module/function.

July 1 -July 7:

- Implement some other OAT.methods i.e. validation, data quality assessment, filtering on time series.

- Debugging.

July 8 -July 14:

- GUI to have OAT.method functionalities i.e. validation, data quality assessment, filtering on time series.
- Testing using istSOS-sensor of methods.

July 15 -July 21:

- Implement remain OAT.methods.
- Testing on these methods.

July 22 -July 29: (Phase 2 Evaluation)

- Testing on istSOS sensors.
- Testing the overall working of each and every module/function.

July 30 -Aug 6:

- Implement CKAN (Open Data Portals) plugin.
- Implement publish datasets ready to be harvested using CKAN.
- Testing and Debugging

Aug 7 -Aug 13:

- Documentation
- Testing
- Results and comparisons of different methods on different datasets.
- Cleaning and Optimizing the codes.

Aug 13 -Aug20: (Pencils down date)

- Take a week to scrub code, write a few test cases and test them, improve the current documentation.
- This is the Firm 'pencils down period. I will test and review all the code I have written by that time.

Aug21 : (final evaluation)

- Buffer period. (Final Week. Prepare documentation for final submission.)
- This is the Firm 'pencils down period. I will test and review all the code I have written by that time.

Do you understand this is a serious commitment, equivalent to a full-time paid summer internship or summer job?
Yes, I understand that this is a serious commitment. I have no exams or other errands during this time. I will be available on IRC and mailing list throughout the summer.

Do you have any known time conflicts during the official coding period?
No.

What is your School and degree?
School: National Institute of Technology, Srinagar (J&K) India.
Degree: 3rd year Computer Science Student (B.Tech.)

Would my application contribute to my ongoing studies/degree? If so, how?
Yes, this project will definitely give me a boost in the sensor management operation domain, observation and analysis. This is related to my last research project field of interest which will help me in SOS field in near future.

Why am I interested in this specific coding project:
I have chosen this project because I have worked on remote sensing project. I want to contribute to the open source development society in sensor field. I want to implement these techniques in an open source software.

General computing experience:

Operating System:	GNU/LINUX, Windows
Programming Languages:	C, C++, Java
Scripting Languages:	Python, Ruby
Internet technologies:	HTML, CSS, Javascript, Node.js
Server Side Scripting:	Mod_python, CGI
Database management System:	MySql, MangoDB

Graphics and UI Development:	OpenGl
Frameworks:	Python(Flask), CakePHP
Other Tools and libraries:	OpenCV, Matlab, Android Studio, Eclipse, Qt

Details of previous GIS experience:
I have not worked on GIS systems before but in my research internship on Remote Sensing, I have some knowledge about it.

Why am I interested in GIS and open source software:
I wanted to know more about GIS. I have some knowledge about it from my course on Remote Sensing. This project is also related to my field (image processing) of interest in which I have done research internship in winter(Nov 20,2016 to Mar 5,2017) at IISC Banglore, India. Moreover, I have earlier contributed to the open source development. It is exciting and challenging in working and contributing to the open source development society.

Why am I interested in working for OSGeo and the software project that I have selected:
This project gives me an opportunity of learning how work is done in the field of SOS(Sensor Observation Service). OSGeo gives me a platform for working with sensor management and statistical tools. I had worked on remote sensing techniques previous winter internship. It is an exciting new area at the intersection of sensor management.

Explain how your SoC task would benefit the OSGeo member project and more generally the OSGeo Foundation as a whole:
Visual analysis invariably plays a very strong role in all aspects of SOS(Sensor Observation Service). The project will develop a whole new range of tools where users can use them in their application and can also be used by researchers in their research work. Moreover no other organizations provide these types of functionalities together.

Do Have you participated to Gsoc before?
No.

Do Have you submitted/will you submit another proposal for this year's GsoC to a different org?
No.

Some of my projects and contribution work are:

- **Data Fusion of remote sensing image(satellite images)**

  I have done research project in data fusion of remote sensing data (Panchromatic and Hyperspectral images) of satellite using java on the time series visualisation analysis of data of satellite at Indian Institute of Science Banglore (India) and pan-sharpening and interpolation of images using Cubic B-spline technique and work on its applications of images(sub-pixel).

  Technical Environment: Java

- I have done project on python(flask), java, android,Qt(c++) and Image Processing project found link [1].

- I had contributed some other open source organisation fossasia(ask_susi), kde(G-comprise) and fixed many bugs [3].

About me:

I started programming when I was in 8th grade (2009) because I was intrigued by how a computer worked. I self-taught myself how to program and now I am proficient with developing web apps (Python and Node.js) and mobile apps for Android platform in various technologies/languages. I have worked with several small businesses as a consultant and have helped them by building customized software solutions to empower their business.

I am an avid hackathon attendee, I love to code entire night and build something cool which solves a real life problem. I am also very fast learner and it helps me work in a fast paced environment and learn new things at a rapid pace compared to my peers including new programming languages/technologies.

I designed some start-up module android and web apps. Which is designed in python (flask) and cakePHP or android development.

I had learnt android development including Java and built a prototype of a location based networking app (Smokey) within a short duration of 1 month. I've open sourced this app and I have been receiving positive messages from people who appreciated so many new things they learnt from reading the source code. This is something which got me motivated to get into open source community and help people. This makes me an excellent resource to learn something new and contribute quickly. I like to build high performance apps and websites and I strive for optimizations which shaves even 1ms in response time.

Reference:

[1] : https://github.com/rahulworld/ [My projects]

[2] : https://github.com/istSOS/android-demo/pull/3 [PR istSOS android-demo]

[3] : https://github.com/istSOS/istsos3-unit-conversion/pull/1 [contribution].