Potential Agenda Items

Thursday, 16 October 2014 10:49 AM

Thursday 23 October 2014

1. Clarify functional requirements for Python system

a. Identify key workflow steps for phase 1 implementation

i. From the workflow diagrams shown in AGDCAppWorkshop\_SambucaRecode, which sections are in scope for the Python system to be developed in phase 1?



1) Just the grey "optimisation module", plus some input data wrangling and writing of output rasters?

b. For each key workflow step:



i. ii.

What are the data inputs? What task is being performed?

1) Algorithm

2) Data wrangling – this needs to be discussed with Simon Oldfield – assume at this stage that we are processing one pixel at a time.

a) format transformation – pixel is a spectrum

b) Tiling

c) Reassembly of tiles

iii. .

What outputs are expected?Range of raster outputs

iv.Testing: (not yet)

1) Known/expected outputs for given inputs

2) Are single-pixel unit tests plausible?

v. Is there only one way to do this step, or would it be reasonable for a user to configure/select/supply an alternate function? Configurable steps would include parameter start ranges and methods and the error evaluation.

1) Example of configurable function is error evaluation

c. What precision is required?

i. ii. iii.

Single

Double yes at least

Fixed to one of the above, or selectable for each run?

d. Are there any visualisation/plotting requirements? no

i. My preference is to leave this out of the Python code. Instead, write the outputs to a standard file format (eg HDF or NetCDF with [CF convention metadata](http://cfconventions.org/)), and leave visualisation to any number of other tools (ArcGIS, QGIS, Grass, Panoply, gdal etc) great – sounds perfect (except for netCDF, we need to discuss with AGDC developers regarding format.

1) *The Climate and Forecast (CF) conventions are metadata conventions for earth science data, intended to*

*promote the processing and sharing of files created with the NetCDF Application*

2) CF and COARDS conventions are supported in ArcGIS 10.1+ (and possibly in 9)

3) I can't confirm directly, but from what I have read, ArcGIS support for HDF is not as good as [support for](http://resources.arcgis.com/en/help/main/10.1/index.html%23/What_is_netCDF_data/004600000001000000/)

[NetCDF](http://resources.arcgis.com/en/help/main/10.1/index.html%23/What_is_netCDF_data/004600000001000000/)

e. PySambuca fully functional both with and without AGDC: what is the best way to achieve this?yes please



i. Staged delivery?

1) AGDC first or last?

first

ii.

Separate Python workflows and wrapper functions?

1) Call into common per-pixel processing methods (we don’t understand)

iii. iv.

Is AGDC data consumption mixed in with other spatial data inputs? Yes -

Another option: AGDC tile assembly as pre-processing step for Sambuca.

No – this is a AGDC function (discuss with AGDC developers)

1) Not making use of AGDC parallel execution

2) Treating AGDC as a data source only

f. AGDCAppWorkshop\_SambucaRecode presentation, slide 5



i. ii. iii. iv.

Project 1: recoding to open source software for implementation on the NCI and drawing from the AGDC.

Project 2: porting to NCI/AGDC - -not porting but building the gui interface on the NCI.

Does this mean I don't need to interface with the AGDC until the second project? No – we do need the data wrangling and interface with the AGDC as part of the first phase, but the gui interface would be phase 2 (discuss luigi and details on govdex as the workflow manager of preference on the NCI).

I thought that AGDC data consumption was part of the first phase, and that the second phase was primarily

concerned with adding the GUI layer.

Yes

g. Is phase one required to operate on a time series, or a single point in time? Time series data are discrete time epochs – time slices not time ranges or steps.

i. For a time series, care must be taken with selection of data from AGDC, as data may be available for a range of times within the span of a single time step (see comment above)

2. Evaluation criteria

a. Performance

i. ii. iii.

For a range of fixed problems, how does performance of PySambuca compare to IDL? Are there any minimum performance targets for the new system?

Or do you just want better scaling and performance than IDL? The faster the better?

b. Validity

i. ii.

Is it possible to validate the final outputs statistically or numerically?

For each function/algorithm, automated unit tests can be created, comparing against expected outputs

1) Either analytically derived, or based on the IDL implementation

iii.

Integration test

1) Run the same input data and configuration through both IDL and Python implementations. Compare the outputs.

a) It is unreasonable to expect numerically identical outputs from IDL and Python due to the many

potential points of divergence

i) EG: different implementation of Amoeba finding slightly different minima

3. Issues:

b) But eyeballing the output images will at least provide a guide.

a. [Numerical Recipes license](http://www.nr.com/licenses/) (Amoeba function) [expressly forbids distribution in source form.](http://www.nr.com/licenses/redistribute.html)



i. ii. iii.

This includes "derivative works", such as your Amoeba implementation

Expressly includes ports to other languages

Suggested responses:

1) Use the [Nelder-Mead downhill simplex minimiser function from SciPy.](http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.optimize.minimize.html)

a) Just like Amoeba, it does not require derivative functions

2) Customisable approach

a) deal with this in the same way as the error function:

i) ii) iii)

Defined interface between workflow code and minimiser function

User configuration supplies the actual function to use

Default option is SciPy Nelder-Mead

4. What lessons can we learn from Luke's work? What problems can we avoid this time?

a. More training and support for Linux HPC workflows b. GUI interface can be less intimidating

c. Parallel execution needs to be easy to configure and use

5. Roll-over into a second project without seeking a new endorsement is possible in specific circumstances:

a. The work being rolled over does not represent new features that are outside the terms of reference of the original

endorsement.

b. An appendix to the original RFP should be prepared for the roll -over request. This needs to indicate the reasons for

the roll-over and what features are expected to be deferred.

c. No free pass. Roll-over proposals are considered in the competitive process of RFP approvals alongside all other

proposals.

d. Although it is likely, it should not be assumed that the rolled -over project will be approved, or that it will be assigned to the same SA team member.