Processing in EPOS

Context of processing in EPOS

Complete EPOS report on Processing available at: https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/42/comments/318/attachments/375/download

Summary of EPOS requirements for processing

Detailed requirements

Inputs

The type of processing inputs depends of the community. For example, Seismology community has: time series, earthquake parameters (metadata), the data is not heterogeneous but it has to be compliant to community standards). Satellite imaging processing is also an important use case. The rate is not real-time processing.

How the data made available to the analytics phase, depends on the community again. For seismologists, the data is available by web-services also stream-based. Others may support common services and transfer protocols (e.g. ftp, gridftp, rsync).

As an example of input data we can consider: Seismology and Satellite communities, Text files, NET-CDF, time-series.

Analytics

EPOS has several uses cases that can be categorized by HPC (Simulation), Data-Intensive. In the case of HPC uses cases, they need over 1000 cores. EPOS can paralyze most of the Data-Intensive uses cases (e.g. misfit pipelines). The time taken for processing can be different depending on the use case. Simulation can have duration from half an hour.

The process implementation depends a lot in each case of community. Once again, we are using the Seismology community information for provide the following information:

- Programming languages: Mostly Python and Fortran.
- Platform (hardware, software): HPC-clusters, MPI, Stream processing, dispel4py, obspy, multiprocessing, Python 7 and Matlab.
- Specific software requirements: Python 7, Matlab, Fortran, R.

EPOS needs to support WPS (work processing services), because it is considered as one of the standards. However, it is not yet implemented in the communities. Mostly relying on interactive job or SaaS platform (VERCE – ESA GEP).

There is the possibility/willingness for scientists and practitioners to inject/execute proprietary/user defined algorithms/processes.

For testing and tune an algorithm/process, some test/sandbox should be used, but it is not defined what/how in EPOS.

Old school scientists use interactive processing, but EPOS is trying to change this behaviour and move to batch processing.

EPOS has a sort of monitoring console based on Provenance.

EPOS sometimes uses/perceives their processes like a black box or a workflow

EPOS uses different workflows for processing, for example: WSP-grade, dispel4py, investigating others. And also reuse sub-process across processes, especially in data acquisition, and in platform control workflows.

EPOS has in place (as examples of processes):

- Forwards modelling, which includes earthquake simulation and misfit analysis.
- Satellite image processing.
- Geohazards, which includes processing for geohazard evaluations
- Anthropogenic hazards (hazard that are not natural per se, but they are provoked by human activities).

More information about processes examples can be found at [2,3,4]

Output

The data outputs can be: Imagines, time series datasets, plots, 3d geometrics and movies (videos). The size of the data depends of the type of the data (from a MB to GB). Regarding with the rate, one of the approaches will be to let users to tune the rate of production.

The data is published through a catalogue exposing information about datasets and collections using CERIF.

Statistical

Data can be collected with an hypothesis in mind and also can be something being measured, but mostly the first one.

Normally, in EPOS, users collect the data, they formulate a question, and then run a process to get the results and evaluate their hypothesis. Later, users can tune their questions, but the data will remain the same (re-use the same data for all their questions/hypothesis).

References:

1. http://www.gaia-clim.eu/
2. http://wiki.services.eoportal.org/tikiindex.php?page=Geohazards+Platform
3. https://portal.verce.eu/home
4. http://is-epos.eu/home,2,en.html

Formalities (who & when)

Go-between	Rosa Filgueira
RI representative	Alessandro Spinuso
Period of requirements collection	From September to November 2015
Status	Finished