|  |
| --- |
| **Requirements survey topics:**   1. General question 2. Identification and citation 3. Curation 4. Cataloguing 5. ***Processing*** 6. Provenance 7. Optimization 8. Community support |

**ENVRIplus Theme 2:**

Requirements information gathering exercise

***ICOS (Integrated Carbon Observation System)***

*RI representative(s):*

* *Ute Karstens,*

*ICOS Carbon Portal & Lund University*

* *Margareta Hellström,*

*ICOS Carbon Portal & Lund University*

* *Dario Papale*

*ICOS Ecosystem Thematic Center (University of Tuscia, Italy)*

* *Benjamin Pfeil,*

*ICOS Ocean Thematic Centre & Geophysical Institute, University of Bergen*

This version is from January 27, 2016.

**4. Processing**

**CASE 1: Atmospheric Thematic Center**

The ATC has chosen not to provide specific answers to this part.

**CASE 2: Ecosystem Thematic Center (Dario Papale, University of Tuscia)**

1. Data processing desiderata: input
   1. What data are to be processed?

We have more than 50 final variables, from images to 10Hz numeric data

***What are their:***

* + 1. Typologies, i.e. “classifications” of the data like tabular data, images, etc.

Images, Samples, tabular data

* + 1. Volume, i.e. what is the “size”. There are many ways to indicate data volume including volume in bytes, volume in number of entries in a dataset, number of dataset to analyse, number of files;

Images are 25Mb each, 40 images per measurement day, 5-8 days per year. Samples are 50 soil and vegetation samples. Tabular data are continuous from 10Hz to 1 minute time resolution, about 200 variables.

* + 1. Velocity, i.e. the “rate” at which data are produced and expected to be analyzed;

From produced 10Hz analyzed daily to samples every 5 years analyzed in 1 month

* + 1. Variety, i.e. how heterogeneous your datasets are. It might happen that the data to be analyzed do not fit with a well-defined relational schema/format or that the data are incomplete, etc.

Incredibly variable for each site, highly standard for the same variable among sites.

* 1. How is the data made available to the analytics phase? By file, by web (stream/protocol), etc.

BADM system, using file/web/app and imported in database

* 1. Please provide concrete examples of data.

Sonic and IRGA data for eddy covariance, meteo variables, soil sampels, digital hemispherical pictures, ceptometer measuremetns, litter weight, diameter of trees…

1. Data processing desiderata: analytics
   1. Computing needs quantification:
      1. How many processes do you need to execute?

Between 10 and 50?

* + 1. How much time does each process take/should take?

Really variable dependent. Some is daily and takes 30-60 minutes, some is periodic and takes days (e.g. chemical analysis).

* + 1. To what extent processing is or can be done in parallel?

Can be done in parallel for sure.

* 1. Process implementation:
     1. What do you use in terms of:
  + *Programming languages?*

Fortran, C++, Python. Some still in Matlab but going to be converted in Python.

* + *Platform (hardware, software)?*

Different workstations

* + *Specific software requirements?*

no

* + 1. What standards need to be supported (e.g. WPS) for each of the known/expected processes?

Cannot answer

* + 1. Is there a possibility/willingness for scientists and practitioners to inject/execute proprietary/user defined algorithms/processes?

Not at the moment

* + 1. Do you use/expect to use a sandbox to test and tune the algorithm/process?

Could be

* 1. Do you use batch or interactive processing?

yes

* 1. Do you use a monitoring console?

Not really a console but we follow the processing and errors. Not sure is this is equivalent

* 1. Do you use/perceive your processes like a black box or a workflow for processing?

I perceive it as a workflow, definitely not a black box. But don’t use any tool

* + 1. If you use a workflow for processing, could you indicate which one (e.g. Taverna, Kepler, proprietary, etc.)

No

* + 1. Do you reuse sub-processes across processes?

yes

* 1. Please provide concrete examples of processes to be supported/currently in use;

Very difficult to answer. We calculate eddy covariance fluxes, look to EddyPro code to have an idea…

1. Data processing desiderata: output
   1. What data are produced?

Different data, not easy to classify

***Please provide:***

* + 1. Typologies, i.e. “classifications” of the data like tabular data, images, etc.

Mainly tabular, some is also provided as image.

* + 1. Volume, i.e. what is the “size”. There are many ways to indicate data volume including volume in bytes, volume in number of entries in a dataset, number of dataset to analyse, number of files;

Half-hourly, 100-500 variables probably. Plus the more sporadic measurements, small volume.

* + 1. Velocity, i.e. the “rate” at which data are produced and expected to be analyzed;

From daily to every 5 years

* + 1. Variety, i.e. how heterogeneous your datasets are. It might happen that the data to be analyzed do not fit with a well-defined relational schema/format or that the data are incomplete, etc.

Mostly tabular so quite heterogeneous.

* 1. How are analytics outcomes made available? Do you expect them to be “automatically published” through a “catalogue”? Do you expect every scientist is provided with a web-based “workspace” where the results are stored?

Not relevant for ETC

1. Statistical questions
   1. Is the data collected with a distinct question/hypothesis in mind? Or is simply something being measured?

Strange question, nothing is collected “just to be collected”. ICOS objective is to make measurements. (Updated 2016-01-27.)

* 1. Will questions/hypotheses be generated or refined (broadened or narrowed in scope) after the data has been collected? (N.B. Such activity would not be good statistical practice)

Not relevant for ETC

1. Statistical data
   1. Does the question involve analyzing the responses of a single set of data (univariate) to other predictor variables or are there multiple response data (bi or multivariate data)?

Multiple data streams used

* 1. Is the data continuous or discrete?

Both (if you consider 10Hz continuous)

* 1. Is the data bounded in some form (i.e. what is the possible range of the data)?

Yes, variable specific

* 1. Typically how many datums approximately are there?

Cannot answer

1. Statistical data analysis
   1. Is it desired to work within a statistics or data mining paradigm? (N.B. the two can and indeed should overlap!)

both

* 1. Is it desired that there is some sort of outlier/anomaly assessment? Which one? Do you have concrete examples that are used in your community/domain?

We have a number of plausibility tests for the data, starting from ranges to advances statistical tests. Impossible to explain here.

* 1. Are you interested in a statistical approach which rejects null hypotheses (frequentist) or generates probable belief in a hypothesis (Bayesian approach) or do you have no real preference?

Not able to answer at the moment, we have tests of both the approaches.

**CASE 3: Ocean Thematic Center (Benjamin Pfeil, University of Bergen)**

1. Data processing desiderata: input
   1. What data are to be processed?

Discrete and underway measurements in the field of marine biogeochemistry

***What are their:***

* + 1. Typologies, i.e. “classifications” of the data like tabular data, images, etc.

Tabular data

* + 1. Volume, i.e. what is the “size”. There are many ways to indicate data volume including volume in bytes, volume in number of entries in a dataset, number of dataset to analyse, number of files;

Up to 5e+6 bytes for a single file, up to 100 files per year for the OTC at the initial stage

* + 1. Velocity, i.e. the “rate” at which data are produced and expected to be analyzed;

Depends on the platform from a single measurements every few seconds up (underway observations) to every few years (discrete observations)

* + 1. Variety, i.e. how heterogeneous your datasets are. It might happen that the data to be analyzed do not fit with a well-defined relational schema/format or that the data are incomplete, etc.

Maybe up to 10 different types by now at this stage but more in the future

* 1. How is the data made available to the analytics phase? By file, by web (stream/protocol), etc.

We are not operational yet and have delayed mode files but from 2016 we will start to stream data.

* 1. Please provide concrete examples of data.

We have ICOS data included within SOCAT ([www.socat.info](http://www.socat.info) - [http://ferret.pmel.noaa.gov/SOCAT\_Data\_Viewer/UI.vm#panelHeaderHidden=false;differences=false;autoContour=false;globalMin=413.2591;globalMax=468.6698;xCATID=socat\_erddap;xDSID=socatV3\_c6c1\_d431\_8194;varid=fCO2\_recommended-socatV3\_c6c1\_d431\_8194;imageSize=auto;over=xy;compute=None;constraintCount=2;constraint0=text\_cr\_none\_cr\_none\_cr\_WOCE\_CO2\_water\_cr\_2\_cr\_WOCE\_CO2\_water\_cr\_2\_cr\_eq\_cr\_;constraint1=variable\_cr\_socatV3\_c6c1\_d431\_8194\_cr\_fCO2\_recommended-socatV3\_c6c1\_d431\_8194\_cr\_fCO2\_recommended\_cr\_NaN\_cr\_fCO2\_recommended\_cr\_NaN\_cr\_ne\_cr\_;constraintPanelIndex=6token;catid=socat\_erddap;dsid=socatV3\_c6c1\_d431\_8194;varid=fCO2\_recommended-socatV3\_c6c1\_d431\_8194;avarcount=0;xlo=-180;xhi=180;ylo=-80;yhi=90;tlo=01-Jan-1957%2000:00;thi=31-Dec-2014%2000:00;operation\_id=Trajectory\_interactive\_plot;view=xyt](http://ferret.pmel.noaa.gov/SOCAT_Data_Viewer/UI.vm#panelHeaderHidden=false;differences=false;autoContour=false;globalMin=413.2591;globalMax=468.6698;xCATID=socat_erddap;xDSID=socatV3_c6c1_d431_8194;varid=fCO2_recommended-socatV3_c6c1_d431_8194;imageSize=auto;over=xy;compute=None;constraintCount=2;constrai) and http://pangaea.de/search?ie=UTF-8&env=All&count=10&q=socat+2015+-code

1. Data processing desiderata: analytics
   1. Computing needs quantification:
      1. How many processes do you need to execute?

Not being done just for ICOS OTC data yet

* + 1. How much time does each process take/should take?

Not being done just for ICOS OTC data yet

* + 1. To what extent processing is or can be done in parallel?

Not being done just for ICOS OTC data yet

* 1. Process implementation:
     1. What do you use in terms of:
  + *Programming languages?*

Python

* + *Platform (hardware, software)?*

Unix

* + *Specific software requirements?*

LAS

* + 1. What standards need to be supported (e.g. WPS) for each of the known/expected processes?

Common ones

* + 1. Is there a possibility/willingness for scientists and practitioners to inject/execute proprietary/user defined algorithms/processes?

Will be implemented for discrete measurements for adjustments based on cross-over analysis.

* + 1. Do you use/expect to use a sandbox to test and tune the algorithm/process?

No answer provided.

* 1. Do you use batch or interactive processing?

Will be most likely implemented for parts of data

* 1. Do you use a monitoring console?

No answer provided.

* 1. Do you use/perceive your processes like a black box or a workflow for processing?

No answer provided.

* + 1. If you use a workflow for processing, could you indicate which one (e.g. Taverna, Kepler, proprietary, etc.)

No answer provided.

* + 1. Do you reuse sub-processes across processes?

No answer provided.

* 1. Please provide concrete examples of processes to be supported/currently in use;

No answer provided.

1. Data processing desiderata: output

Looks similar to the answers above.

* 1. What data are produced?

See 2) above

***Please provide:***

* + 1. Typologies, i.e. “classifications” of the data like tabular data, images, etc.

See 2) above

* + 1. Volume, i.e. what is the “size”. There are many ways to indicate data volume including volume in bytes, volume in number of entries in a dataset, number of dataset to analyse, number of files;

See 2) above

* + 1. Velocity, i.e. the “rate” at which data are produced and expected to be analyzed;

See 2) above

* + 1. Variety, i.e. how heterogeneous your datasets are. It might happen that the data to be analyzed do not fit with a well-defined relational schema/format or that the data are incomplete, etc.

See 2) above

* 1. How are analytics outcomes made available? Do you expect them to be “automatically published” through a “catalogue”? Do you expect every scientist is provided with a web-based “workspace” where the results are stored?

Centrally stored and published

1. Statistical questions
   1. Is the data collected with a distinct question/hypothesis in mind? Or is simply something being measured?

Tricky question: we try of course to measure in areas that are relevant for CO2 uptake or release – so we simply measure there and to answer scientific questions. The combination of all measurements is needed to verify or deny hypothesis

* 1. Will questions/hypotheses be generated or refined (broadened or narrowed in scope) after the data has been collected? (N.B. Such activity would not be good statistical practice)

No

1. Statistical data
   1. Does the question involve analyzing the responses of a single set of data (univariate) to other predictor variables or are there multiple response data (bi or multivariate data)?

The latter

* 1. Is the data continuous or discrete?

Both

* 1. Is the data bounded in some form (i.e. what is the possible range of the data)?

Limited by time and space…didn’t get the question

* 1. Typically how many datums approximately are there?

Up to 30.000 lines with up to 12 columns

1. Statistical data analysis
   1. Is it desired to work within a statistics or data mining paradigm? (N.B. the two can and indeed should overlap!)

No answer provided.

* 1. Is it desired that there is some sort of outlier/anomaly assessment? Which one? Do you have concrete examples that are used in your community/domain?

No answer provided.

* 1. Are you interested in a statistical approach which rejects null hypotheses (frequentist) or generates probable belief in a hypothesis (Bayesian approach) or do you have no real preference?

No answer provided.

**CASE 4: Atmospheric Inverse Modelling Systems (Ute Karstens, Lund University)**

Answers in this section are based on the requirements of atmospheric inverse modelling systems. These are run to estimate net greenhouse gas (GHG) fluxes between the surface and the atmosphere from an optimal fit to atmospheric GHG concentration measurements, usually including prior constraints on the flux estimates. In these inversion schemes, atmospheric transport models are implemented to relate concentrations in the air with fluxes from and to the Earth’s surface.

1. Data processing desiderata: input
   1. What data are to be processed?

* Measurements of atmospheric trace gases (e.g. CO2, CH4) that are available from different networks such as ICOS(Europe) and NOAA (global), including measurement uncertainty estimates
* A-priori estimates of the spatio-temporal distribution of trace gas fluxes (e.g. biosphere fluxes from more or less sophisticated vegetation models) and anthropogenic emission inventories, including uncertainty estimates
* Three-dimensional time-dependent data on atmospheric variables (wind, temperature, pressure) provided by meteorological centers like ECMWF. These represent meteorological forcings.

***What are their:***

* + 1. Typologies, i.e. “classifications” of the data like tabular data, images, etc.

Different types of data: tabular data (time series), 3-4 dimensional\* global fields (model output, inventories)

\*Dimensions of the fields are: geographical latitude, longitude, time and eventually height above the earth surface

* + 1. Volume, i.e. what is the “size”. There are many ways to indicate data volume including volume in bytes, volume in number of entries in a dataset, number of dataset to analyse, number of files;

Depending on data types: ~ 1GB (time series), ~2TB (fields)

* + 1. Velocity, i.e. the “rate” at which data are produced and expected to be analyzed;

6 monthly - yearly

* + 1. Variety, i.e. how heterogeneous your datasets are. It might happen that the data to be analyzed do not fit with a well-defined relational schema/format or that the data are incomplete, etc.

Different types of datasets (time series, 3-4D fields) but internally consistent

* 1. How is the data made available to the analytics phase? By file, by web (stream/protocol), etc.

Web portals, ftp transfer of files

* 1. Please provide concrete examples of data.

Greenhouse gas concentration time series, emission inventories, meteorological reanalysis

1. Data processing desiderata: analytics

Atmospheric transport models can be global or regional, based on either eulerian or lagrangian approaches. They are run in “forward mode” to compute concentrations based on fluxes and in “inverse mode” to compute the response function of a particular region to a given measurement. Inverse procedures combine concentration measurements, a priori fluxes and transport (including the particular uncertainties) in a statistical way to produce an optimized estimate of surface fluxes (incl. their uncertainties).

The inversion typically consists of the following steps:

1. Calculation of prior trace gas concentration using the atmospheric transport model together with the prior flux estimates.

2. Minimization of the cost function to derive posterior fluxes. The cost function combines the difference matrix between measured and model simulated concentration (icl. uncertainties) and the difference matrix between prior and posterior fluxes (incl. uncertainties). This is usually done in an iterative way, requiring ca. 50-100 iterations (forward + inverse transport model runs).

* 1. Computing needs quantification:
     1. How many processes do you need to execute?

Depends on the definition of process, but main processes are preparation of input data, atmospheric transport model runs, optimization, analysis of model output.

* + 1. How much time does each process take/should take?

The main process (transport model ruins + optimization) needs approx. 50000 CPU hours, corresponding to approx. 10 days real time on a HPC system with parallel processing.

* + 1. To what extent processing is or can be done in parallel?

Mostly consecutive steps, but parallel processing within the steps

* 1. Process implementation:
     1. What do you use in terms of:
  + *Programming languages?*

Fortran, Python, C++, R

* + *Platform (hardware, software)?*

HCP computers, Unix or Linux operating systems

OpenMPI, MPI, netCDF, HDF, GRIB, compilers (e.g. Fortran, C)

* + *Specific software requirements?*

Libraries for scientific computing (e.g. openMPI, MPI, netCDF), visualization and analysis software (e.g. python, NCL, GrADS, IDL, netCDF utilities)

* + 1. What standards need to be supported (e.g. WPS) for each of the known/expected processes?

NetCDF CF conventions

* + 1. Is there a possibility/willingness for scientists and practitioners to inject/execute proprietary/user defined algorithms/processes?

Yes, most of the analysis tools are user defined.

* + 1. Do you use/expect to use a sandbox to test and tune the algorithm/process?

The use of a sandbox environment would be desirable

* 1. Do you use batch or interactive processing?

batch systems

* 1. Do you use a monitoring console?

The individual steps of the processing chain are monitored by checking intermediate output files and log files.

* 1. Do you use/perceive your processes like a black box or a workflow for processing?
     1. If you use a workflow for processing, could you indicate which one (e.g. Taverna, Kepler, proprietary, etc.)

Specific workflow monitoring systems are not yet often used but should be explored.

* + 1. Do you reuse sub-processes across processes?

Most computer programs are modular so that subroutines can be used in different parts of the processing.

* 1. Please provide concrete examples of processes to be supported/currently in use;

Optimization techniques, e.g. Kalman filter or conjugate gradient methods

1. Data processing desiderata: output
   1. What data are produced?

* Optimized spatio-temporal distribution of surface fluxes incl. uncertainty estimates.
* Posterior (best-estimate) 3-D concentration fields derived from a transport model run based on the optimized fluxes.

***Please provide:***

* + 1. Typologies, i.e. “classifications” of the data like tabular data, images, etc.

3-4 dimensional fields, time series

* + 1. Volume, i.e. what is the “size”. There are many ways to indicate data volume including volume in bytes, volume in number of entries in a dataset, number of dataset to analyse, number of files;

1-2 TB

* + 1. Velocity, i.e. the “rate” at which data are produced and expected to be analyzed;

6 monthly - yearly

* + 1. Variety, i.e. how heterogeneous your datasets are. It might happen that the data to be analyzed do not fit with a well-defined relational schema/format or that the data are incomplete, etc.

Complete 3-4 dimensional fields, time series

* 1. How are analytics outcomes made available? Do you expect them to be “automatically published” through a “catalogue”? Do you expect every scientist is provided with a web-based “workspace” where the results are stored?

Results will be stored on the computing system and made available at a data portal

1. Statistical questions
   1. Is the data collected with a distinct question/hypothesis in mind? Or is simply something being measured?

The purpose of the data is defined before starting the measurements.

* 1. Will questions/hypotheses be generated or refined (broadened or narrowed in scope) after the data has been collected? (N.B. Such activity would not be good statistical practice)

No, the initial questions are not redefined, but new questions/hypothesis may arise

1. Statistical data
   1. Does the question involve analyzing the responses of a single set of data (univariate) to other predictor variables or are there multiple response data (bi or multivariate data)?

Multivariate data

* 1. Is the data continuous or discrete?

Most of the data sets are continuous

* 1. Is the data bounded in some form (i.e. what is the possible range of the data)?

Most of the data is a measure of natural processes and therefore bounded within their natural limits.

* 1. Typically how many datums approximately are there?

Time series: ~106, 3-4D fields: >1012 but not all datums are statistically independent

1. Statistical data analysis
   1. Is it desired to work within a statistics or data mining paradigm? (N.B. the two can and indeed should overlap!)

The approaches are based on statistics and large datasets, both paradigms apply to some extend.

* 1. Is it desired that there is some sort of outlier/anomaly assessment? Which one? Do you have concrete examples that are used in your community/domain?

Anomalies are analysed within the system, but obvious outliers, e.g. due to instrument failure, are removed before ingesting the data.

* 1. Are you interested in a statistical approach which rejects null hypotheses (frequentist) or generates probable belief in a hypothesis (Bayesian approach) or do you have no real preference?

The analyses are often based on a Bayesian approach.