Processing requirements

Introduction

Data Processing or Analytics is an extensive domain including any activity or process that performs a series of actions on dataset(s) to distil information [Bordawekar 2014]. It is particularly important in scientific domains especially with the advent of the 4th Paradigm and the availability of "big data" [Hey 2009]. It may be applicable at any stage in the data life cycle from QA and event recognition close to data acquisition to transformations and visualisations to suit decision makers as results are presented. Data analytics methods draw on multiple disciplines including statistics, quantitative analysis, data mining, and machine learning. Very often these methods require compute-intensive infrastructures to produce their results in a suitable time, because of the data to be processed (e.g., huge in volume or heterogeneity) and/or because of the complexity of the algorithm/model to be elaborated/projected. Moreover, these methods being devised to analyse dataset(s) and produce other "data"/information (than can be considered a dataset) are strongly characterised by the "typologies" of their inputs and outputs. In some data-intensive cases, the data handling (access, transport, IO and preparation) can be a critical factor in achieving results within acceptable costs.

In fact, when analysing the needs of Research Infrastructures involved in ENVRIplus we focused on collecting four major aspects that characterise each RI's data processing needs:

- **Input**, i.e., what are the characteristics of the dataset(s) to be processed? This includes dataset(s) typologies, volume, velocity, variety /heterogeneity, and access methods;
- **Analytics**, i.e., what are the characteristics of the processing tasks to be enacted? This includes computing needs quantification, implementation aspects including programming languages, standards and re-use potential;
- **Output**, i.e., what are the characteristics of the products resulting from the processing? This includes typologies, volume, variety, variety /heterogeneity, and availability practices;
- **Statistics**, i.e., what are the scientific motivations leading to the identification of the specific data processing envisaged by a community. This includes aspects related to data collection and hypothesis generation.

Each of these are summarised below.

Overview and summary of processing requirements

Input

As largely expected, RIs' needs with respect to dataset(s) to be processed are quite diverse because of the diversity in the datasets that they deal with. Dataset(s) and related practices are diverse both across RIs and within the same RI. For instance, in EPOS there are many communities each having its specific typologies of data and methodologies (e.g., FTP) and formats (e.g., NetCDF, text) for making them available. Time series and tabular data are two very commonly reported types of dataset to be processed yet they are quite abstract. In what concerns "volume", dataset(s) vary from a few KBs to GBs and TBs. In the large majority of cases dataset(s) are made available as files while few infrastructures have plans to make or are making their data available through OGC services, e.g., ACTRIS.

The need to homogenise and promote state-of-the-art practices for data description, discovery and access is of paramount importance to provide RIs with a data processing environment that makes it possible to easily analyse dataset(s) across the boundaries of RI domains.

Analytics

When moving to the pure processing part, it emerged that RIs are at diverse levels of development and that there is a large heterogeneity. For instance, the programming languages currently in use by the RIs range from Python, Matlab and R to C, C++, Java, and Fortran. The processing platforms range from the 3 Linux servers in the case of ACTRIS to HPC approaches exploited in EPOS. No major issues emerged with respect to licences. Software in use or produced tends to be open source and freely available. In the majority of cases there is almost no shared or organised approach to make available the data processing tools systematically both within the RI and outside the RI. One possibility suggested by some RIs is to rely on OGC/WPS for publishing data processing facilities.

Some care needs to be taken balancing the benefits of common solutions with the need to support a wide range of working practices well – we return to this in Section 4.2. The platform should be "open" and "flexible" enough to allow

1. scientists to easily plug-in and experiment with their algorithms and methods without bothering with the computing platform,
2. service managers to configure the platform to exploit diverse computing infrastructures,
3. third-party service providers to programmatically invoke the analytics methods, and
4. to support scientists executing existing analytic tasks eventually customising/tuning some parameters without requiring them to install any technology or software.

Output

In essence, we can observe that the same variety characterising the input is there for the output also. I this case, however, it is less well understood that there is a need to make these data available in a systematic way, including information on the entire process leading to the resulting data. In the case of EMBRC it was reported that the results of a processing task are to be made available via a paper while for EPOS it was reported that the dataset(s) are to be published via a shared catalogue describing them by relying on the CERIF metadata format.

In many cases, but by no means all, output resulting from a data processing task should be "published" to be compliant with Open Science practices. A data processing platform capable of satisfying the needs of scientists involved in RIs should offer an easy to use approach for having access to the datasets that result from a data processing task together. As far as possible it should automatically supply the entire set of metadata characterising the task, e.g., through the provenance framework. This would enable scientists to properly interpret the results and reduce the effort needed to prepare for curation. In cases where aspects of the information are sensitive, could jeopardise privacy, or have applications that require a period of confidentiality, the appropriate protection should be provided.

Statistical

Only a minority of the RIs within ENVRIplus responded to the statistics questions within the processing requirements gathering. We know from the ENVRI project that LifeWatch had the support of a wide range of statistical investigations, not just biodiversity, as part of its mission. Unsurprisingly given the diversity of the component RIs, there were a variety of different attitudes to the statistical aspects of data collection and analysis. One RI (IS-ENES-2) felt that data analysis (as opposed to collection) was not their primary mission whereas for others (e.g., within EMBRC researchers at the University of St Andrews) reaching conclusions from data is very much their primary purpose.

As environmental data collection is the primary aim of many of the RIs it appears that day-to-day consideration of potential hypotheses underlying data collection is not undertaken. Hypothesis generation and testing is for scientific users of the data and could take many forms. However, some RIs (e.g., LTER and ICOS) stressed that general hypotheses were considered when the data collection programmes and instruments were being designed especially if the data fed into specific projects. Hypotheses could be generated after the fact by users after data collection and indeed this would be norm if data collection is primary a service to the wider scientific community.

RIs can be collecting multiple streams of data often as time series, thus there is the potential to undertake multivariate analysis of the data. Again unsurprisingly given the diversity of science missions, there was no consistency in approaches. Data could be continuous and discrete, be bounded by its very nature or have bounds enforced after collection. Data sets are potentially very voluminous; total data sets with billions of sample points might be generated. Most analysers will be engaging in formal testing of hypotheses rather than data mining although the latter was not necessarily ruled out. Many RIs had or are going to implement outlier or anomaly detection on their data.

Again unsurprisingly given the potential uses for the data, a variety of statistical methods can be undertaken. RIs did not feel restricted to working solely within either a frequentist or Bayesian framework. Much of the data collected takes the form of time series.

The current mission of ENVRIplus will address the aspects of data collection, preparation and integration that should provide a context for such statistical approaches. The integration of tools and statistical methods, and their mapping onto platforms, should be supported in an appropriate virtual research environment or science gateway. This requires collaborative R&D building on experience from the EU project Biodiversity Virtual eLaboratory (BioVeL). This would fully integrate statistical analysis tools with the data handling, and map the processing tasks automatically to appropriate data-intensive subsystems and computational resources. The sustainable path, which would also promote international exchanges of environmental-data analysis methods, would benefit from collaboration with organisations such as the NSF-funded Science Gateway Institute. This environmental-analytical virtual eLaboratory kit is a good example of a candidate common subsystem, where the balance of a core used by many RI communities with tailoring to support specialised working practices would need careful investigation. Providing such an integrated combination of data lifecycle support with easily activated and steered analysis and visualisation tools will improve researcher productivity by removing many hurdles they have to get over today. This will accelerate discovery and evidence production, but it will also boost those who take those results and present them to decision makers. This will interact with the arrangements for federation support –see Section 4.2.3.

Research Infrastructures

The following RIs contributed to developing processing requirements

<Delete from the following list any that were not able to contribute on this topic>

<Add an interest inducing sentence or two, to persuade readers to look at the contribution by a particular RI. e.g., What aspect of the summary of requirements, or the special cases, came from this RI. Check with RIs that they feel they are correctly presented.>

ACTRIS: *<e.g., This RI ... and therefore has XYZ <Topic> requirements, with a particular empahsis on ...>*

AnaEE:

EISCAT-3D:

ELIXIR:

EMBRC:

EMSO:

EPOS:

Euro-ARGO:

EUROFLEETS2:

ESONET:

EUROGOOS:

FIXO3:

IAGOS:

ICOS:

INTERACT:

IS-ENES2:

JERICO:

LTER:

SEADATANET:

SIOS: