Explaining eco-informatics and the ecological data chain

by Avinash Chuntharputsat, information management scientist, SAEON

Data management can be a daunting task especially to those new to the field of eco-informatics.

To start this process a definition of eco-informatics needs to be looked at. The following is an explanation from Wikipedia:

“Eco-informatics, or ecological informatics, is the science of information (informatics) in ecology and environmental science. It integrates environmental and information sciences to define entities and natural processes with language common to both humans and computers. However, this is a rapidly developing area in ecology and there are alternative perspectives on what constitutes eco-informatics” [1].

A few definitions have been circulating, mostly centered on the creation of tools to access and analyse natural system data. However, the scope and aims of eco-informatics are certainly broader than the development of pedestrian metadata standards to be used in documenting datasets. Eco-informatics aims to facilitate environmental research and management by developing ways to access and integrate databases of environmental information, and to develop new algorithms enabling different environmental datasets to be combined to test ecological hypotheses.

Eco-informatics characterise the semantics of natural system knowledge. For this reason, much of today’s eco-informatics research relates to the branch of computer science known as “knowledge representation, and active eco-informatics projects are developing links to activities such as the Semantic Web [1].”

Keeping this explanation in mind, it emerges that eco-informatics deals with the complexities of handling ecological and environmental data. Fig. 1 is a flowchart of the ecological data handling process.

Fig. 1: Flowchart of the ecological data handling process.
The reasons for this being that the complexity of the data process chain involves the planning of experiments and observations, methodology used, stringency in collecting data, accuracy of electronically capturing data and the dexterity of preserving data. Inclusive in the data chain process is the analysis and reuse of data to yield higher level information products and the accuracy and relevancy of the analysis.

From Fig. 1, three important components emerge from the data chain. These are the metadata, quality control and quality assurance. Metadata is the data/information about the data. A good metadata record is needed along each step of the data chain. The entire history of the development of the dataset should be recorded and tracked in the metadata. A suitable standard for the capture and storage of metadata should be used. The Ecological Metadata Language (EML) is one such standard that can be used for ecological metadata. More information on EML and metadata can be accessed from the following sites:

- www.ecoinformatics.org and

The other two components are quality control (QC) and quality assurance (QA). In the diagram, QA and QC are associated with different steps of the data chain. Quality assurance is largely associated with the data creation (such as planning and methodology) and analysis steps, while quality control is largely associated with the data capturing and archiving steps. For a better understanding of why and how this occurs, a closer look at the definitions of QA and QC are needed.

For the purposes of this exercise, the definitions of QA and QC used by the Intergovernmental Panel on Climate Change (IPCC) are found to be suitable. These definitions are based on greenhouse gas emissions but can be applied in a broader ecological context. The definitions from Chapter 8 of the "IPCC Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories" are as follows:

“Quality control (QC) is a system of routine technical activities, to measure and control the quality of the inventory as it is being developed. The QC system is designed to:

(i) Provide routine and consistent checks to ensure data integrity, correctness, and completeness;
(ii) Identify and address errors and omissions;
(iii) Document and archive inventory material and record all QC activities.

QC activities include general methods such as accuracy checks on data acquisition and calculations and the use of approved standardised procedures for (emission) calculations, measurements, estimating uncertainties, archiving information and reporting. Higher tier QC activities include technical reviews of source categories, activity and (emission) factor data, and methods.

Quality assurance (QA) activities include a planned system of review procedures conducted by personnel not directly involved in the inventory compilation/development process. Reviews, preferably by independent third parties, should be performed upon a finalised inventory following the implementation of QC procedures. Reviews verify that data quality objectives were met, ensure that the inventory represents the best possible estimates of emissions and sinks given the current state of scientific knowledge and data available, and support the effectiveness of the QC programme” [3].

For the implementation of QC activities various software packages are available for particular fields of research. These packages aid in the identification of missing values and other errors such as statistically significant outliers. However, a strong understanding of the nature of the data is a prerequisite for quality control activities.

In the definition of QA, relevance is made to reviews by a third party. This is particularly relevant to the SAEON situation. The SAEON nodes which conduct research and observations in different bioclimatic regions of the country have node liason committees. These committees are convenient to act as independent auditors of data quality. Many data producing organisations have such technical committees. These committees could also play an important role in ensuring that the data is of the highest standards for their relevant organisations.

References