New technologies for automating the management and analysis of conservation data

by Judith Kruger (SANParks), Matt Jones (NCEAS), Mark Schildhauer (NCEAS) and Nicola Stevens (SAEON)

South African National Parks (SANParks) encompasses many different ecosystems, reflecting the tremendous diversity of life forms and landscapes that exist in South Africa. The intelligent management of these precious resources is a critical goal for SANPark managers and scientists alike. Accurate, comprehensive and timely data are needed to inform decisions, and the analyses themselves will often require sophisticated computing equipment and software to model and transform the data in a multi-step process. However, the development of complex ecological and ecosystems analyses currently presents a serious challenge to scientific researchers and conservation decision makers, due to the difficulties in accessing the appropriate data, and using these to inform well-defined, replicable analyses.

Current data management and analytical approaches are not well-suited for accommodating the often heterogeneous and distributed data that typify integrative, site-based science [1]. Various types of data are typically collected and archived in custom databases created by independent researchers, so these are not connected to one another, and the data are largely inaccessible for cross-cutting analysis and modeling. Even after a potentially arduous and time-consuming exercise in accessing and integrating the relevant data for some scientific research question, the analyses themselves often require multi-step modelling and transformation of the data. This analytical process is typically not easily replicated, so that after weeks or months of effort, one might have some interesting results, but would be unable to readily replicate, share, or modify their models as new or corrected data, or alternative analyses are suggested.

This article describes some emerging technologies in databases and analysis software which are assisting researchers and managers in improving the adaptive management approaches to conservation science within SANParks. We describe recent advances in technology that focus on facilitating better access to and analysis of a broad range of conservation data within SANParks. These solutions are based on a standard for metadata, EML, that comprehensively describes both the structure and contents of arbitrary, heterogeneous data resources. We describe a production-ready database framework for storing and presenting these metadata across the SANParks network, based on free, open source software. Finally, this metadata framework is closely integrated with a scientific workflow application, Kepler, which allows researchers to actively document their analyses such that various data resources on the network are clearly identified, as well as the step-by-step process of modelling and analysing the data. The Kepler application also allows a researcher to readily share analyses with their colleagues and others, such that the results can be readily re-executed, and/or replicated using new or modified data.

Documenting heterogeneous data

One solution to the problem of storing and documenting data is to use a flexible, structured metadata standard such as Ecological Metadata Language (EML). EML is a metadata standard developed by the ecological community in the USA primarily to be used by ecologists [2]. The standard evolved over time incorporating various elements from other standards and is supported by a wide range of users ensuring that the standard is continually improved. EML can be used for documentation of both metadata and datasets, it is modular and very extensible so it can be used to describe many different data formats. It can also be exploited by a number of computer applications. EML was developed to address the lack of dataset documentation as well as to provide structure to traditionally unstructured information [see 3]. EML is both human and machine readable and will assist in the development of long term archives.

EML allows us to specify multiple levels of documentation. The first level is for identification of the datasets and this is documented in the "title", "abstract", and "keyword" fields. The next level is the discovery level information. This includes information on both the geographic extent of the datasets and well as the temporal coverage. The geographic extent is captured with spatial coordinates which can then be used to plot coordinates on a map. Applications can leverage both the geographical and date time fields for dataset discovery. The evaluation level metadata includes information about the methods as well as the project level information. The methods indicate how the data was collected and may be described as a set of hierarchial processes with substeps. The access level information indicates who may change and read the data and metadata and also describes where the data can be obtained and the format of the data. The final level is the logical model information which describes the structure of data tables and their variables. Each variable needs a name, description, and a measurement type.

Once the EML has been created for a particular data package which includes the metadata and the data tables, it can be uploaded into a data repository where it can be browsed, downloaded and archived. The Knowledge Network for Biocomplexity (KNB) now has 16 partner organisations who contribute EML data and metadata to a global data repository. The advantages of being part of this global system is that the data is replicated between the sites so this means that the risk of losing data is minimal. The access rules do not change with replication which means that the dataset owner can still determine who gets access to the data without being concerned that they will be amended once replication occurs.

Simplifying metadata entry

Gathering quality metadata about heterogeneous datasets can be time consuming and therefore difficult without supporting software tools. SANParks uses the free software product Morpho [4] to create EML metadata which is then uploaded into the SANParks data repository (http://dataknp.sanparks.org). Morpho creates EML using a wizard front end for the simpler metadata and an EML tree for the more complex metadata that is not used a lot. We have used Morpho to upload data to the SANParks data repository, which currently contains many ecological datasets from the Kruger National Park and only a few datasets from the other national parks. With time this system will be expanded to include all the ecological data from the other national parks. An extension to this system allows spatial layers with their FGDC metadata that is created through ArcGIS to be incorporated into the repository. This means that spatial and non-spatial data can be searched together and downloaded if needed.

The South African Environmental Observation Network (SAEON) aims to detect, predict and react to environmental change. The collection of reliable, well-documented, long-term data goes a long way in achieving this goal. This work is actively in motion through the archiving of environmental datasets, the collection of additional data through the implementation of monitoring programs and the analysis of existing long-term datasets. This will then provide reliable, long-term data that can subsequently inform scientific research and decision making in the context of global environmental change.

These activities are very data intensive and have resulted in a collection of many different datasets. Data ranges from that collected by members of the public to data collected by SAEON and SAEON-associated students. All of which were collected for a range of different purposes using various methodologies.

SAEON Ndlovu Node needed a data management system that could incorporate such a range of datasets and has settled on the use of Ecological Metadata Language to document our environmental datasets. Like SANParks, SAEON uses the free software, Morpho, to create EML metadata which is then uploaded onto an online registry (http://dataknp.sanparks.org). The Ndlovu Node has recently started using this data management system and has documented approximately 20 datasets for the savanna biome (outside national parks) and approximately five datasets for the Arid Zone Node.

In line with the SAEON data policy, SAEON encourages the public sharing of data and metadata through the online data registry; however some data received by SAEON requires that access to it be restricted e.g. data not yet published. Therefore the only requirement for archiving a dataset with SAEON, on the online registry, is that all datasets must be accompanied by metadata. The metadata is recorded, using EML, by the SAEON Ndlovu Node data manager in communication with the data provider. If the data provider requires restricted access, only the metadata will be publicly viewable. The data provider receives a login account and can view the full dataset online and can grant permission for certain users if so desired. The SAEON data policy also requires that data collected by SAEON through monitoring and research or using SAEON resources be made available online free-of-charge, with as few restrictions as possible. This data is being archived every six months by the data manager.

Using EML to create metadata has definite advantages in that it not only provides good descriptions of SANParks datasets in ways that are useful to researchers and managers, but it does so using a format that is increasingly being used by ecologists at parks and

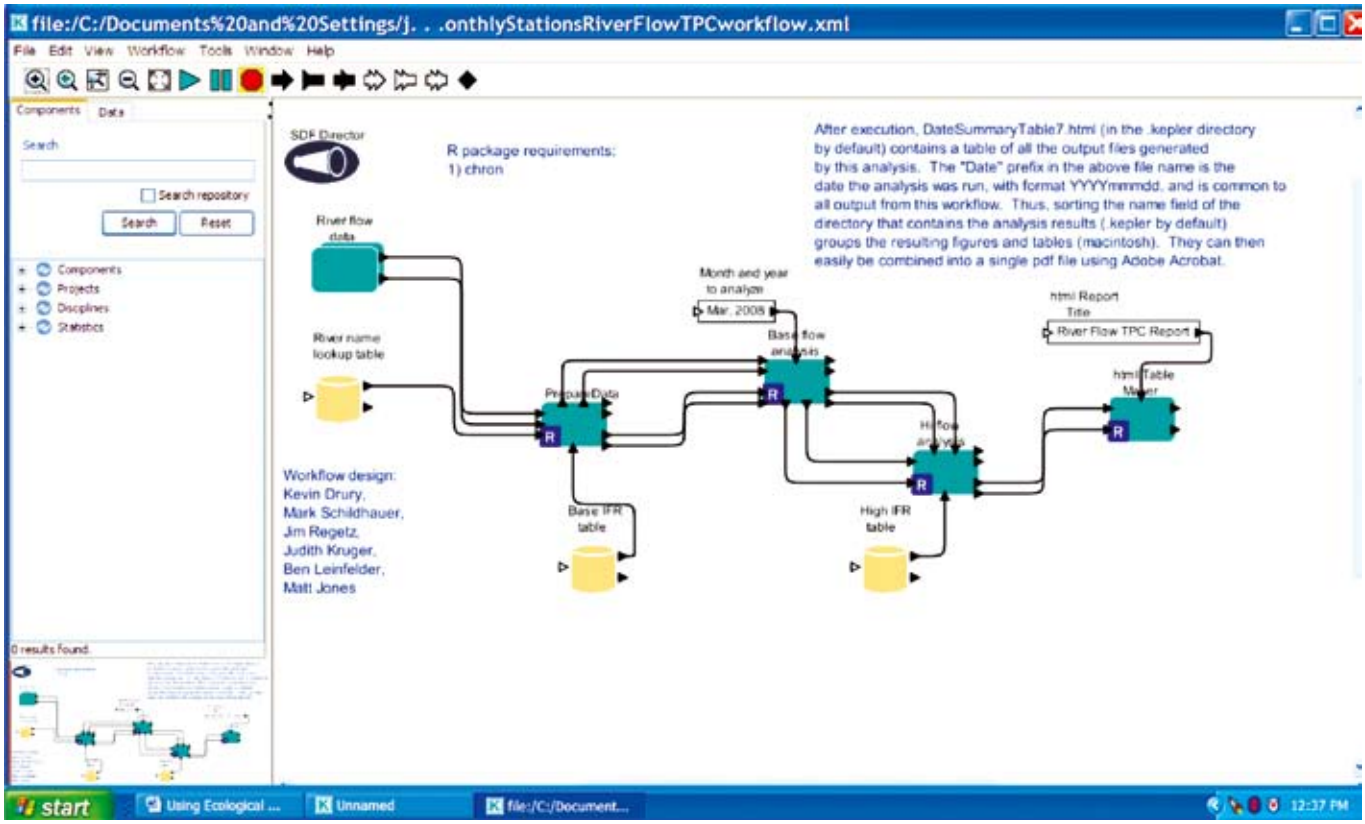

Fig. 1: The Kepler scientific workflow application, showing a workflow that processes river-flow data and performs an analysis to detect whether current conditions have exceeded the established thresholds for concern and management action in the park.

field stations around the world. It is hoped that as more ecological communities in South Africa become aware of these technologies and start taking part, that the network will widen and sharing of ecological data in South Africa will increase.The usefulness of the SANParks data repository will grow as more researchers routinely store their metadata and data in it. We believe this data repository not only provides the best guarantee that precious scientific information about SANParks will not become lost, but also that scientists and managers will have increasingly exciting and useful ways of directly interacting with the data repository holdings. We next describe one such application that can powerfully use EML metadata for accomplishing scientific analyses.

Metadata-driven analysis using scientific workflows

Once metadata are in such a data repository, software tools can access this information and utilise the metadata to start understanding the data. The network of data repositories allows users to search for data from any of the participating sites using keywords. An example of this would be: if one was interested in above-ground biomass production, a keyword search for this would locate all the data that adheres to the search criteria irrespective of where this data was collected and who was the contributing site. This means that one can start integrating global ecological data allowing for cross-scale or even trans-continental analysis.

We are applying the data stored in the SANParks data repository in standardised analyses to detect exceedances in certain ecological variables. These exceedances or thresholds of potential concern (TPCs) are undesirable levels set for certain ecological variables. An example of this is the concern about the river levels being too low and a lower level of acceptability or an instream flow requirement (IFR) has been determined for each river. The monitoring data is analysed for each river in order to determine whether these lower limits or IFRs have been exceeded or not.

To accomplish these cross-scale and cross-discipline analyses, one needs an analytical tool that can interface with a variety of data sources like the SANParks and KNB data repositories and that provides access to commonly used analytical tools. One emerging approach to this is to use a scientific workflow system such as Kepler [5] to orchestrate the analysis. Scientific workflows in general, and Kepler in particular, have the capability to access multiple analytical systems such as R Matlab, and other statistical and modelling systems. Many scientific analyses need to access features of different software packages as a multi-step process, and Kepler provides a way to integrate these, so that the analysis is ultimately well-organised and documented, as well as executable. In this sense, Kepler "orchestrates" the analyses, by providing an application in which different software packages can be linked together for execution.

We have extended Kepler to be able to directly access data and metadata that are stored in the SANParks and KNB data repository.

Each workflow can consist of a series of analytical steps starting from the data cleaning step and finishing with a complex analytical procedure to produce outputs (Fig. 1). The advantages of using a workflow approach is that it is always clear what version of a dataset was used in a particular analysis and one can then rerun any analysis on the same or different data. The workflows and the analytical outputs can be saved to the data repository.

SANParks has deployed Kepler as the analytical tool to analyse the data collected by their monitoring programs (Fig. 1). Fig. 1 shows a workflow that uses river-flow data to determine if the agreed upon amount of water or IFR has been met. Fig. 2 shows the results of the river-flow TPC analysis with critical periods when the TPC was exceeded (below the IFR) flagged in red on the graph. In these cases, the river-flow metadata and data are stored in the SANParks data repository, and available over the internet. The workflow can access the latest versions of these data, and rapidly and repeatedly generate updated analyses and visualisations of trends in the data.

Currently Kepler is a desktop application that produces outputs on the local computer on which it is running. This approach is useful for the analyst developing the workflow, and even for some scientists running workflows for their own purposes. However, park managers need these analyses to be run on a regular basis and need the outputs to be shared broadly with a variety of management stakeholders. Thus, we are developing a web-based execution and reporting system that allows outputs from workflow runs to be directed to a web page, thereby enabling managers to access and use them in making informed management decisions while only needing a simple web browser (e.g. Firefox or Internet Explorer) of their choice.

Conservation scientists and managers must rely on comprehensive and timely access to accurate information about the

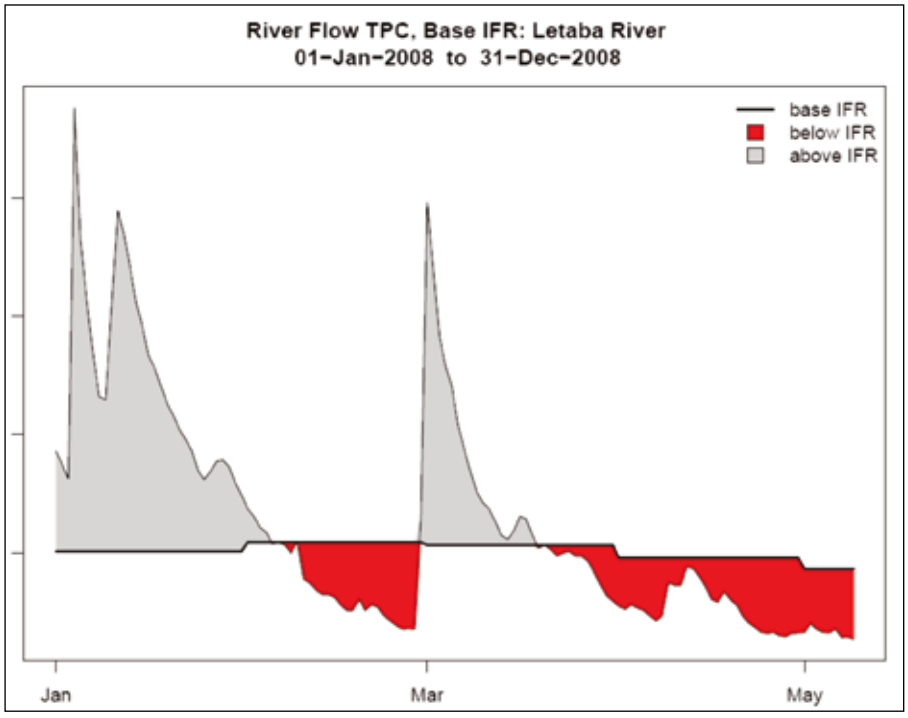

Fig. 2: River-flow TPC analyses for the Letaba River.

land holdings under their stewardship, in order to make critical decisions regarding how to effectively sustain the precious biodiversity and ecosystem services under their charge. This is not a straightforward challenge, for conservation and ecological analyses typically require gathering together diverse types of data, and modelling and transforming these using multi-step, potentially sophisticated techniques. SANParks personnel need a simple, consistent way in which to share both their data and analyses in effective ways, for meeting this challenge.

This paper describes a framework that we are successfully deploying at Kruger National Park, and increasingly more broadly within SANParks, to dramatically facilitate access to and analysis of ecological and environmental data. The framework uses a growing international standard for metadata, EML, for the description of ecological datasets. The power of this framework is further enhanced by its linkage to the scientific workflow application, Kepler, which provides a robust way to document and flexibly share analyses that automatically link to the latest (or desired versions) of monitoring data. We believe these approaches will significantly enhance the capability and efficiency of researchers and managers in monitoring and analysing park data, as well as accelerate the capability of both scientists and other stakeholders to understand the processes underlying the complex ecosystems found within the SAN Parks

References

[1] MB Jones, M Schildhauer, OJ Reichman, and S Bowers: The new bioinformatics: integrating ecological data from the gene to the biosphere, Annual Review of Ecology, Evolution, and Systematics. 2006. Vol. 37, pp 519–544, 2006.

[2] E Fegraus, SJ Andelman, MB Jones and M Schildhauer: Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation, Bulletin of the Ecological Society of America, Vol. 86(3), pp 158-168, 2005.

[3] WK Michener, JW Brunt, JJ Helly, TB Kirchner, SG Stafford: Non-geospatial metadata for the ecological sciences, Ecol. Appl. Vol. 7, pp 330-342, 1997.

[4] D Higgins, C Berkley, and MB Jones: Managing Heterogeneous Ecological Data sing Morpho, Proceedings of the 14th International Conference on Scientific and Statistical Database Management, July 24-26, 2002. J Kennedy (ed), ISBN 0-7695-1632-7 ISSN 1099-3371, 2002.

[5] B Ludäscher, I Altintas, C Berkley, D Higgins, E Jaeger-Frank, M Jones, E Lee, J Tao, Y Zhao: Scientific Workflow Management and the Kepler System. Special Issue: Workflow in Grid Systems, Concurrency and Computation: Practice & Experience, Vol. 18(10), pp 1039-65, 2006.

Heard at the SAEON Summit

- *"South Africa, like most other nations, faces major environmental challenges and unless we have the necessary environmental observations systems that allow us to monitor and fix these, we are in deep trouble."* Dr Bob Scholes, CSIR
- *"No data belongs to you exclusively. It is suicidal to restrict access to data."* Dr Bob Scholes, CSIR
- *"It has become suicidal for an individual, an organisation or a country, to resist joint research projects or to restrict access to important datasets and information, as this would have serious negative consequences for its scientific, environmental and economic competitiveness in the global arena."* Derek Hanekom, Deputy Minister of Science and Technology
- *"We need armies of data managers."* Coleen Moloney, University of Cape Town
- *"Your data centre will be remembered for its data quality only; do one thing to compromise it and nobody will want to touch it."* Laurie Barwell, CSIR
- *"We have a choice where the globe will end up in a hundred years."* Guy Midgley, South African National Biodiversity Institute