NATURE | TOOLBOX: Q&A

My digital toolbox: Ecologist Ethan White on interactive notebooks

IPython project packs text, code, figures and tables into one document.

Richard Van Noorden

30 September 2014

Utah State University

Ethan White uses large data-sets to test ecological models.

Ethan White, an ecologist at Utah State University in Logan, discusses the software and tools that he finds most useful in his research.

How would you describe your research?
I use data generated by literature mining, government surveys and citizen science to understand ecological systems on large spatial scales, at which many crucial ecological processes and environmental problems operate. For example, I try to understand why some regions of Earth have high biodiversity whereas others are relatively species-poor.

Which software or online tools do you use on a regular basis, and why?
I learned about the IPython notebook in early 2012, and was immediately hooked. The first time I opened one up, it was clear that this tool was going to change the way I worked. I have been using it for both teaching and research ever since. The IPython notebook is a free tool that lets you combine formatted text, code, and the figures and tables that code generates, in a single document.

It allows you to quickly develop a rich scientific document that conducts an analysis, shows the results and explains the scientific context. Notebooks are also easy to share on the web, where others can modify or explore the analysis. IPython notebooks are very easy to learn, work in most web browsers and support a variety of programming languages used for data analysis, including Julia, Python and R (hence the recent name change to Project Jupyter). You do need to know how to program in one of the supported languages to get the most out of them, however.

I have a personal interest in talking about the EcoData Retriever, which is a tool that Ben Morris (a former undergraduate in my research group) and I built. It makes working with the broad array of ecological data sets faster, easier and more reproducible. Data come in all shapes and sizes, and often need a lot of checking, formatting and cleaning before they can be used to do science. I was frustrated with the amount of time my group was spending on data cleaning and database set-up, so we built a tool to make our lives easier.

The EcoData Retriever automatically finds, downloads, cleans up and installs an array of ecological data sets in a number of common formats (such as Microsoft Access, SQL or text files) for analysis. The Retriever needs instructions for handling new data sets, but this is usually as simple as a few lines of plain text.

Visit the Toolbox hub for more articles

In my lab, we use Git and GitHub to manage all of our code, much of our data and even some of our paper writing. Git is a version-control system that lets you keep track of changes to documents — one of the first computational best practices that I implemented in my research group. GitHub is a web-based hosting service, built on Git, that makes it easy to collaborate and share research. (For more on these tools, see nuclear engineer Katy Huff on version-control software.)

Which emerging tools do you have your eyes on?
I think the dat project by Max Ogden (a programmer based in Oakland, California) is really intriguing. His team is building tools that allow data to be continuously synchronized among all of the major data-base and file-storage systems. This has the potential to make it much easier to work with the diverse array of data that are out there.

What software would you like to see developed in future?
Many of the most interesting questions in science require combining data from numerous different sources. We need software that will help to automate this process, so that scientists can spend more time doing science, and less time cleaning up and merging data.

Nature | doi:10.1038/nature.2014.16015