(In?)Extricable Links between Data and Visualization: Preliminary Results from the VISTAS Project

Judith Cushing[1], Evan Hayduk[1], Jerilyn Walley[1], Lee Zeman[1],
Kirsten Winters[2], Mike Bailey[2], John Bolte[2], Barbara Bond[2], Denise Lach[2],
Christoph Thomas[2], Susan Stafford[3,] Nik Stevenson-Molnar[4]

[1]The Evergreen State College, [2]Oregon State University, [3]University of Minnesota,
[4]Conservation Biology Institute (Corvallis OR)
judyc@evergreen.edu

1 Introduction

Our initial survey of visualization tools for environmental science applications identified sophisticated tools such as *The Visualization and Analysis Platform for Ocean, Atmosphere, and Solar Researchers* (VAPOR) [http://www.vapor.ucar.edu], and *Man computer Interactive Data Access System* (McIDAS) and *The Integrated Data Viewer* (IDV) [http://www.unidata.ucar.edu/software]. A second survey of ours (32,279 figures in 1,298 articles published between July and December 2011 in 9 environmental science (ES) journals) suggests a gap between extant visualization tools and what scientists actually use; the vast majority of published ES visualizations are statistical graphs, presenting evidence to colleagues in respective subdisciplines. Based on informal, qualitative interviews with collaborators, and communication with scientists at conferences such as AGU and ESA, we hypothesize that visualizations of natural phenomena that differ significantly from what we found in the journals would positively impact scientists' ability to tune models, intuit testable hypotheses, and communicate results. If using more sophisticated visualizations is potentially so desirable, why don't environmental scientists use the available tools?

We suggest two barriers to using sophisticated scientific visualization: lack of the desired visualizations, addressed elsewhere [1, 2], and difficulty preparing data for visualization, addressed here. As David Maier remarks in his Keynote Address to this conference: Big data [3] "implies a big variety of data sources, e.g., multiple kinds of sensors…on diverse platforms…coming in at different rates over various spatial scales…. Few individuals know the complete range of data holdings, much less their structures and how they may be accessed" [4]. In this paper, we identify two major data issues that scientists face relevant to their use of visualization tools: (1) the complexity of their own data and (2) complex input data descriptors for visualization software and perceived (or actual) difficulty of transforming data prior to using those tools.

We briefly describe our own project (VISTAS), articulate our collaborators' data structures, and report on data requirements for a subset of visualization software identified as useful for ES. We conclude that the complexity of input data formats is so daunting for most scientists and even for visualization researchers and developers,

adfa, p. 1, 2011.

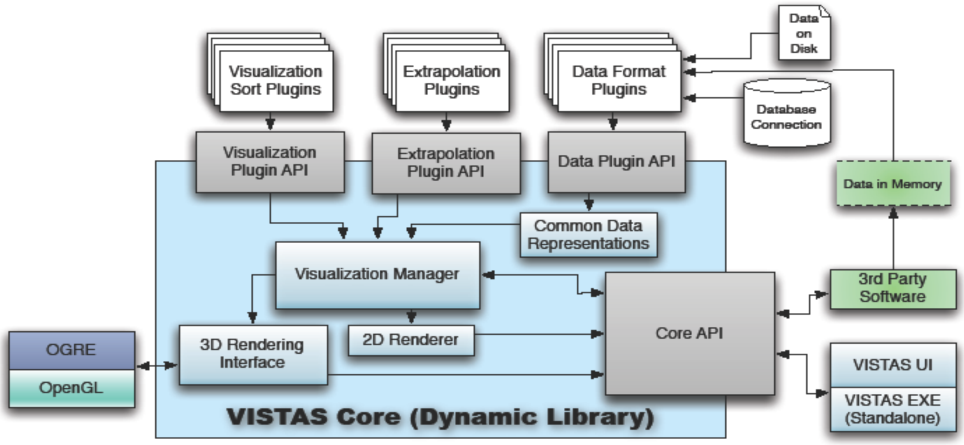
that visualizations that might advance science, and would certainly advance communication of research results by scientists to colleagues and the public, remain undone. We thus strongly encourage the scientific database community to address the problems scientists face in characterizing and transforming their data.

2 VISTAS Project Overview

Our prior work suggested that visual analytics can help scientists more effectively use large data sets and models to understand and communicate complex phenomena. We hypothesized that cross-scale visualization of natural phenomena would enable scientists to better deal with massive data stores and understand ecological processes. Visualization of ecosystem states, processes, and flows across topographically complex landscapes should enhance scientists' comprehension of relationships among processes and ecosystem services, and the posing of testable hypotheses [5]. Scientific visualization is not new and much excellent work exists, but few tools easily integrate complex topography with visualizing diverse data [6, 7]. Fewer still allow viewers to scale up or down in space and time, critical to ES grand challenges [8].

The recently National Science Foundation funded VISTAS (VISualization of Terrestrial-Aquatic Systems) project aims to develop and test visualizations so scientists better understand and communicate ES. Objectives include: 1) research visualization needs for our science collaborators and develop a proof of concept tool that meets those needs, 2) conduct ES research with the tool, 3) use social science methods to document development processes and to assess the visualizations, answering the questions: which visualizations are most effective, for what purposes, and with which audiences. VISTAS' ES research focuses on one geographical area west of the crest of the Oregon Cascades, on 3D representations of land use, and on process-based models that simulate cycling and transport of water and nutrients, problems similar to other ES grand challenges.

3 VISTAS Architecture and Data Structures



Here, we briefly describe VISTAS' architecture and data structures, and our close collaborators' data structures. VISTAS' design (above Figure) makes it easy to add new input formats and visualizations. We separated the front- and back-ends, so our scientists could drive their models using visualizations. VISTAS currently uses a raster data model (grids), and does not convert coordinate systems, so height and elevation data must be input along with variable values for a given cell in the grid (in the same row). Future use of OpenDX [http://www.opendx.org] should alleviate that restriction.

VISTAS' close collaborators include Bob McKane and Allen Brookes who model the ecohydrology of watersheds and basins with **VELMA** [9]; John Bolte, author of a land use model **ENVISION**, a GIS-based decision support tool integrating scenarios, decision rules, ecological models, and evaluation indices [10]; and Christoph Thomas who collects spatially distributed point-measurements of air flow, air temperature and humidity using **SODAR** across the landscape, scales from 10s to 100s of meters [11].

VELMA outputs a tiff file per simulation year, two.csv files with daily and annual results, and files with spatial results, in ESRI Grid ASCII format with a single value for each grid cell. File size depends on the number of variables, currently generating 38.5mb/day (10 gb/yr), but in the future to 5x variables, i.e., 50 gb/yr (64 km2). VISTAS now handles VELMA output of 30m cells, 64 km2, 70,000 cells, 100,000 rows, 30 variables, converting csv files to a 2D contiguous grid, speeding processing 10-20x. Elevation is incorporated into each cell, avoiding the need for alignment.

ENVISION stores data as C++ objects, map data as a table, each column a variable, each row mapping to a polygon, and can read NetCDF and rasterize output data. Time steps are yearly. Temporal data stored as delta arrays (map changes) are large, e.g., 10 million elements and designate which polygon changed, start and end values, when change occurred. The ArcGIS-like shape files (points, lines, polygons, a vector data model) are visualized as 2D maps. VISTAS will display ENVISION 3D terrain models, initially 30,000 km2, 180,000 polygons, 10-20 hectares/polygon, a 90m grid.

The **Metek SODAR** sensor [http://www.metek.de] transmits an acoustic pulse at a specified frequency,–then listens for a return signal; data are analyzed to determine wind speed and direction and the turbulent character of the atmosphere. Measurements are compiled into a 42x136 matrix per 10-min. increments. Each file contains one 24-hour period (39,168 records). Metek provides simple visualization software, where users can select a custom time period and defined parameters for measured values (spectrum, potential temperature, inversion height, wind direction or speed), but VISTAS will provide capability of overlaying SODAR with VELMA modeled data.

4 Conclusions

This paper concludes with a synopsis of the two previously cited visualization options for environmental science, and barriers scientists face in using such systems – the same barriers we (and other scientific visualization developers) face as we develop and document input data plug-ins for VISTAS.

Tailored for the astro- and geo-sciences, VAPOR provides interactive 3D visualization on UNIX and Windows systems with 3D graphics cards, handling terascale data. VAPOR can directly import WRF-ARW output data with no data conversion, but data must be on the same grid with the same level of nesting. Full access to features is available if users convert data to VAPOR's format (VDC), but VAPOR provides tools to convert common data formats.

McIDAS is an open source tool for visualizing multi- and hyper-spectral satellite data, and handles many data formats, including Unidata's IDV and VisAD; satellite images in AREA, AIRS, HDF, and KLM; and meteorological data in McIDAS-MS, netCDF, or text; and gridded data in NCEP, ECMWF or GRID formats.

With so many data formats supported by these and other sophisticated packages, why don't more environmental scientists use them? Not surprisingly, these sophisticated tools – even if data transformations are provided – require scientists to understand both their own data structures as well as the tool's input data structure requirements. Our survey of information managers at Long Term Ecological Research sites [http://www.lternet.edu] concluded that even for simple ArcGIS or MatLab visualizations, such effort is beyond the time or expertise of many scientists. Alternatives might be to send data to a visualization center, or establish collaborations with visualization specialists. This works for some scientists, with the funds and time, to produce visualizations for a particular purpose, but our collaborators want to use visualizations interactively – to steer their computational models, help intuit new hypotheses, and explain results to collaborators and other stakeholders.

Without software that characterizes data, scientists need to understand arcane formats and many will eschew valuable tools – and developers of tools like VISTAS, VAPOR and McIDAS will continue to spend time and money writing idiosyncratic data transformations. What Howe et al provide for long tail science [12] is a first step towards removing the inextricable links between data syntax and semantics that will enable scientists to use the tools they need. We encourage the SSDBM community to research and develop data descriptor and transformation tools that separate the characterization and transformation of data from the process of creating semantically meaningful analyses and visualizations.

5 Acknowledgements

The VISTAS project is supported by the National Science Foundation/ BIO/DBI 1062572. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank the anonymous reviewers whose comments helped sharpen the focus of this preliminary report of VISTAS' findings.

References

1. Cushing, J.B. et al: What you see is what you get? Data Visualization Options for Environmental Scientists. Ecological Informatics Management Conference (2011).
2. Schultz, N., and M. Bailey: Using extruded volumes to visualize time-series datasets. Expanding the Frontiers of Visual Analytics and Visualization, J. Dill et al (eds), Springer, 127-148 (2012), ISBN 978-1-4471-2803-8.
3. Alexander, F. et al: Big Data. IEEE Computing in Science & Engineering.13,10-13 (2011).
4. Maier, D.: Navigating Oceans of Data. Abstracts of the Conference on Scientific and Statistical Database Management. http://cgi.di.uoa.gr/~ssdbm12/keynote1.html.
5. Cushing, J.B., et al: Enabling the Dialogue-Scientist<>Resource-Manager<> Stakeholder: Visual Analytics as Boundary Objects. IEEE Intelligent Systems, 24, 75-79 (2009).
6. Smelik, R.M., et al: Survey of Procedural Methods for Terrain Modeling. In: A. Egges, et al (eds) Proc. of the CASA Workshop on 3D Advanced Media in Gaming and Simulation (3AMIGAS). Amsterdam, the Netherlands: 25-24 (2009).
7. Thomas, C.K., et al: Seasonal Hydrology Explains Interannual and Seasonal Variation in Carbon and Water Exchange in a Semiarid Mature Ponderosa Pine Forest in Central Oregon. Journal of Geophysical Research. 114, G04006 (2009).
8. Kratz, T.K., et al: Ecological Variability in Space and Time: Insights Gained from the US LTER Program. BioScience, 53, 57–67 (2003).
9. McKane, R., et al: integrated eco-hydrologic modeling framework for assessing effects of interacting stressors on multiple ecosystem services. ESA Annual Meeting (August 2010).
10. Bolte, J.P., et al : Modeling Biocomplexity - Actors, Landscapes and Alternative Futures. Environmental Modelling & Software. 22, 570–579 (2007).
11. Turner, D.P., Ritts, W.D., Wharton S., Thomas, C.: Assessing FPAR Source and Parameter Optimization Scheme in Application of a Diagnostic Carbon Flux Model. Remote Sensing of Environment, 113.7, 1529–1539 (2009).
12. Howe, B., et al: Database-as-a-Service for Long-Tail Science in Cushing, J., French, J., Bowers, S., (eds) LNCS: SSDBM, 6809, 480-489 (2011).