Editors: Claudio Silva, csilva@cs.utah.edu
Joel E. Tohline, tohline@rouge.phys.lsu.edu

SCIENTIFIC EXPLORATION IN THE ERA OF OCEAN OBSERVATORIES

By António Baptista, Bill Howe, Juliana Freire, David Maier, and Cláudio T. Silva

The authors introduce an ocean observatory, offer a vision of observatory-enabled scientific exploration, and discuss the requirements and approaches for generating provenance-aware products in such environments.

Society's critical and urgent need to better understand the world's oceans is amply documented and has led to a unique convergence of operational and scientific interests in the US, organized around the concept of *ocean observatories*: cyber-facilitated integrations of observations, simulations, and stakeholders. In particular, programs are emerging aimed at creating an operational Integrated Ocean Observing System (IOOS)[1] to address broad society needs and an open, ocean-observing research infrastructure (the Ocean Observatories Initiative [OOI]).[2]

Perhaps no part of the ocean is in more need of observatories than coastal margins, which are among the most densely populated and developed regions in the world. Coastal margins sustain highly productive ecosystems and resources, are sensitive to many scales of variability, and play an important role in global elemental cycles. But natural events and human activities place stresses on these margins, rendering the development of sustainable resources and ecosystems challenging and contentious, with policy decisions often based on insufficient scientific understanding of the causes and consequences of natural and anthropogenic impacts.

Effective scientific exploration thus requires the ability to generate a wide variety of analyses for a broad audience in an ad hoc manner. In this article, we introduce an observatory in evolution, offer a futuristic vision of observatory-enabled scientific exploration, and discuss how provenance is essential to make this vision a reality.

An Observatory in Evolution

CORIE (for *Co*lumbia *Ri*ver *E*stuary) is a coastal margin observatory operated at Oregon Health & Science University since 1996 as an end-to-end, data-to-stakeholder system. In 2004, CORIE became a founding contributing node to the Northwest Association of Networked Ocean Observing Systems (NANOOS), a regional association of IOOS. Since 2006, CORIE has anchored the development of an emerging next-generation coastal-margin observatory—the Science and Technology University Research Network (SATURN)—that forms the core infrastructure of the new Science and Technology Center (STC) for Coastal Margin Observation and Prediction (CMOP), one of 17 STCs funded by the US National Science Foundation across all its areas of activity.

Designed to serve multiple stakeholders[3–5] in Oregon's Columbia River estuary-plume-shelf system, CORIE includes a field observation network and a modeling system with hindcast and forecast capabilities integrated through a sophisticated cyberinfrastructure for delivering data and products on the Web. The Columbia River provides 70 percent of freshwater inflow to the northeastern Pacific Ocean between San Francisco Bay and the Strait of Juan de Fuca. The Columbia River estuary and plume are tightly coupled regions, and both are influenced significantly by changes in the tide, river discharge, ocean conditions, and shelf winds. The plume extends north to British Columbia and south to California, and is a major feature in the upwelling-dominated Oregon–Washington shelf. In the estuary, tides are strong; wetting and drying is extensive; and channel circulation is strong and highly stratified, leading to formation of spatially and temporally variable, ecologically important features such as fronts, internal waves, and estuarine turbidity maxima.

CORIE Observation Network

The observation network has 16 stations in the estuary and two in the plume/shelf. Sensors vary from station to station, but all of them focus on physical properties—for example, they measure temperature at all stations and water level, salinity, and velocity profiles at selected stations. Modes of sensor deployment in stations vary, with buoys used offshore and piles or underwater structures used in the estuary. Most stations are deployed continuously, although their

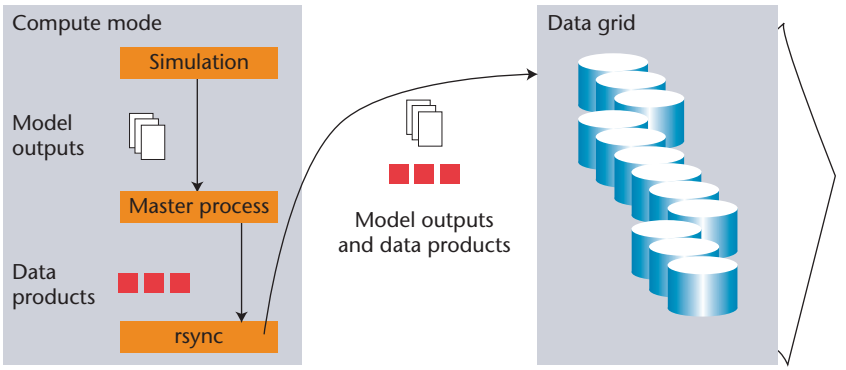

Figure 1. Forecasts. A "master process" script generates data products incrementally and copies the results back to the data grid for dissemination via the Web.

sensor composition can vary over time. The two plume/shelf stations are only deployed spring through fall to limit operational risks during major storms. A real-time telemetry network based on spread-spectrum radios supports both fixed-station and vessel-based measurement. A local vessel equipped with oceanographic instrumentation (salinity, temperature, turbidity, dissolved oxygen, and profiles of velocity) serves as a mobile station with real-time telemetry.

CORIE has an open-access policy for observational data, making it available graphically and for download on the Web in real time (www.stccmop.org/datamart) for public download once it passes an offline quality control step; download of raw data prior to this quality control is facilitated upon request.

CORIE Modeling System

The modeling system generates quality-controlled simulations of cross-scale baroclinic hydrodynamic circulation in the form of daily forecasts, multiyear databases, and process/scenario simulations. All simulations provide 4D (3D space and time) representations of water levels, velocities, temperature, and salinity. In this context, *cross-scale* means the system represents features in the Columbia River estuary and plume and in the continental shelf between northern California and southern British Columbia. Numerical grids are unstructured in the horizontal plane, with the highest resolution in the estuary (roughly 100 meters) and near-plume (roughly 250 meters); typical time steps are 15 minutes. Integral to the system design is automated access to all external forcing (tides, ocean conditions, atmospheric conditions, and river inputs) and to a core set of observations from CORIE and other regional in situ networks.

CORIE uses two unstructured-grid numerical models—Eulerian-Lagrangian *Circ*ulation (ELCIRC)[6] and the Semi-implicit Eulerian-Lagrangian Finite Element (SELFE)[7]—interchangeably as computational engines. Both use semi-implicit Eulerian-Lagrangian numerical methods to solve shallow-water hydrodynamic equations. Their differences lie in the underlying numerical methods (finite volumes with a finite difference approximation of derivatives[6] versus finite elements with finite volumes for the vertical momentum equation[7]) and in the type of coordinate stretching used to represent the domain's vertical axis.

CORIE Cyber-Infrastructure

The information infrastructure consists of a central data grid with four external interfaces: the observation pipeline for acquiring data from remote sensors, the forecast factory for managing daily operational forecasts and generating core data products, the hindcast repository for archiving long-term continuous runs, and an integrative Web site for open data dissemination.

The data grid is a collection of file, database, and Web servers centrally maintained, monitored and managed as one logical unit by CORIE staff. Each machine in the CORIE network is equipped to access the data grid in a uniform way. The observation pipeline manages the streams of data collected at the observation network by performing five crucial processing tasks: acquisition, metadata attachment, transmission, parsing, and registration. Specifically, the system equips data acquired from sensors with metadata and deposits them, unparsed, at a base station. Relational databases provide transactional guarantees and fault tolerance as the data move to a staging database. From there, configurable ingest pipelines process and register the data with the data grid for use by various internal and third-party applications.

The forecast factory's size and scope has grown steadily since its inception in 1997. It currently consists of three Columbia River forecasts supplemented by forecasts for nine other Pacific Northwest estuaries that respond to the same ocean/climate forcing system and forecasts for three estuaries and bays across the world. To accommodate this increased scale, the scheduling, execution, and result processing must be carefully optimized. We've found the factory metaphor particularly useful for modeling this workload:[8] forecasts are allocated to specific compute nodes using scripts that "stage in" all needed input files, launch the workflow, and "stage out"

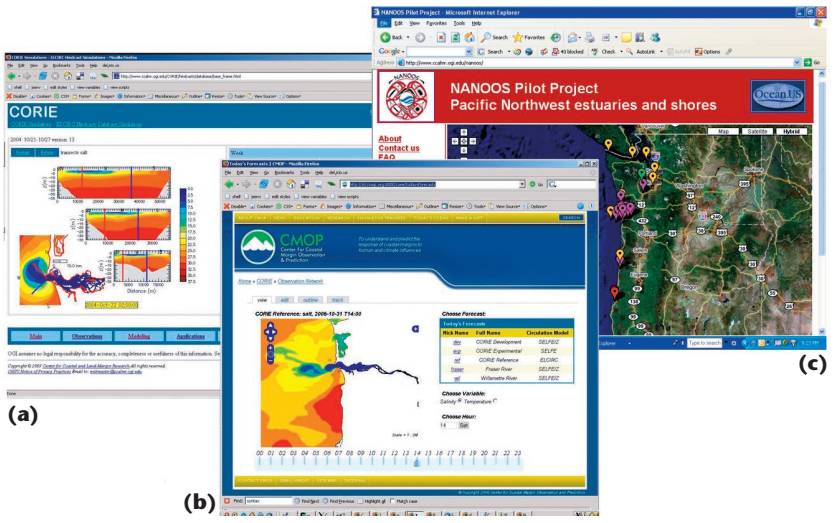

Figure 2. Web applications powered by the CORIE information infrastructure. (a) A browsing interface for the hindcast repository, (b) an application for dynamically generating products from today's forecasts, and the Northwest Association of Networked Ocean Observing Systems (NANOOS) application, which integrates data from various sources.

Figure 3. RoboCMOP. A multisensorial interface could advance observatory-based scientific exploration.

data products to the data grid (see Figure 1). When a new forecast is added, existing forecasts must shift to different nodes to optimize the overall daily workload's completion time. Due to the difficulty of estimating forecast running times at different nodes, administrators use a tool called ForeMan to aid this decision-making process.[8]

Although the emphasis in the forecast factory is reliable daily execution, the emphasis in the hindcast repository is to provide complete and consistent long-term simulations. Latency isn't crucial (as it naturally is with forecasts), but performance is still paramount. Without careful planning, a year of simulated time can take many months to produce on a single processor. Consequently, hindcast runs are typically formulated and executed manually to ensure maximum oversight. Besides fast processors, the crucial technology powering the hindcast repository is a 40-terabyte shared filesystem.

The integrative Web site includes products and services servers to seamlessly disseminate CORIE data along with data from partner institutions. The umbrella site supports a variety of specialized applications; Figure 2 shows three. The technology driving these applications is all free or open source, and much is based on third-party technology.

We predict that many transformative applications for CORIE data and products will be conceived and implemented not by CORIE staff or scientists but by interested third parties with diverse goals in science, education, and public policy. We thus view our task as facilitating such applications by providing simple, powerful, and comprehensive services for accessing, transforming, and visualizing the inventory of observations and model results. Our goal is to empower students, scientists with a wide range of skills, and the general public to construct novel applications using observatory data in minutes or hours rather than weeks or months.

From CORIE to SATURN

As we mentioned earlier, SATURN is an emerging end-to-end observatory

that includes and extensively leverages CORIE's assets and experiences but substantially expands its capabilities. For example, CORIE focuses on the Columbia River estuary and plume, but Saturn explicitly recognizes that the understanding of an estuary/plume system requires a geographically diverse approach that accounts for broader ocean and atmospheric scales. CORIE also focuses on physical variables and processes at fixed stations, whereas Saturn addresses a broad cross-spectrum of physical, chemical, and biological variables and processes adaptively across space and time scales.

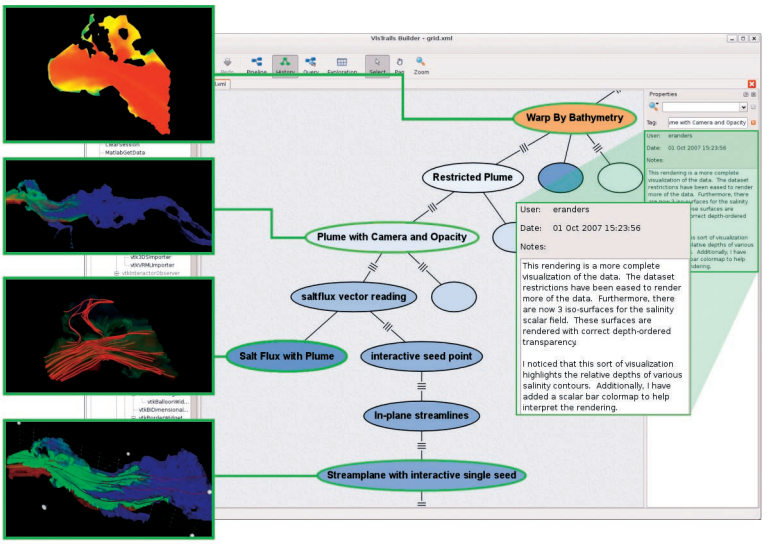

Figure 4. VisTrails.[9] The provenance management system exposes incremental changes to a visualization of model results to users. This provenance history (coupled with explicit annotations as in the inset) lets even nonexperts build new products "by example" instead "from scratch."

A Vision for Scientific Exploration

Future scientific exploration is likely to involve groups that are occasionally geographically distributed and often diverse in expertise. The disparity of expertise in handling and interpreting complex scientific data will be even wider when comparing trained scientists with managers and policy makers who will attempt to use observatories to inform their decisions or with students whose education will depend crucially on unfettered access to observatory data and products.

A major challenge (and opportunity) is thus to facilitate a redefined scientific exploration of ocean data, in which we no longer expect that expert scientists who collect or generate the data sets will also conduct the first line of data analysis. Instead, analysts will face an abundance of heterogeneous data and tools, and they will lack expert knowledge of at least some of these ingredients. Under these circumstances, it will be necessary to expertly assist these analysts. It's useful, as an abstraction, to conceptualize that such assistance will be provided in part by a multi-sensorial software-and-data environment that we refer to as RoboCMOP, as Figure 3 shows. RoboCMOP could advance scientific exploration by accelerating the cycle of science and education, fostering creative thinking, reducing opportunities for key data going unnoticed, and providing tools for capturing, managing, and reusing abstract representations of scientific expertise.

We're currently exploring and extending the provenance management infrastructure provided in the VisTrails system[9] as an example of a platform for managing such representations. The VisTrails system allows the specification of scientific workflows and, more importantly, records any and all adjustments made to a workflow. The result is a thorough provenance history that represents the workflow's evolution, complete with tools to manage, query, and reuse it. Figure 4 shows an example of incremental changes made to a visualization workflow. Each node in the tree represents a workflow version, each one producing a different product. Users can annotate each tree node with descriptive metadata (inset) for another form of provenance.

Not all data and products are specified and generated in a safe, managed, provenance-aware environment such as VisTrails, however. As scientific investigation becomes an increasingly distributed activity, it's likely that we will have to tolerate and leverage unfamiliar information—data sets and products that originated elsewhere and lack complete and accurate provenance.[10]

Informed by the dataspace abstraction[11] that recognizes the need for services to manage heterogeneous and complicated data landscapes, we're also experimenting with a platform called Quarry,[12] which allows incremental accretion of provenance information in the form of metadata attached to data and products rather than to the processes that create them. Quarry doesn't make assumptions about the metadata's form or content, which lets the system accommodate data from arbitrary sources. The system is designed to scale to millions of resources and tens of millions of metadata descriptors by deriving storage structures from patterns mined automatically from the metadata itself. This paradigm is useful in

several situations: when dealing with "other people's data," managing "one-off" products developed outside of a managed environment, exchanging "unfinished" products with third parties, and so on.

Provenance Requirements for Observatories

Let's consider the provenance requirements implied by a RoboCMOP-like environment.We anticipate that these or similar requirements will ultimately apply across ocean observatories and will greatly assist scientific exploration of observatory data and products in science, education, and management.

- *Reusable, active provenance.* An important requirement is the need for tools to explore and re-use provenance information. To enable end-users to actually use provenance and perform their own explorations, or to create new data products on the fly, we need to remove the "person in the middle" (that is, expert programmers). One way to accomplish this is to adopt the model used by social Web sites and Web-based communities (such as Flickr, Facebook, Yahoo! Pipes, or IBM Many Eyes) and develop tools to enable "social analysis of scientific data." The goal is to facilitate collaboration and sharing among users, not only of data but also of analyses. Shared analysis pipelines (workflows) and provenance repositories can expose scientists to computational tasks that provide examples of sophisticated uses for tools. They can also uncover common pipeline patterns that they can re-use to solve different problems. By querying the information in these shared repositories, scientists can leverage the wisdom of the crowds to learn by example; expedite their scientific training; and potentially reduce their time to insight. But for this to become reality, we need to give scientists appropriate and usable tools to explore data in these shared repositories.
- *Pay-as-you-go provenance.* Provenance information must often be collected post hoc; data transformations don't always operate in a controlled environment or can be specially instrumented to emit provenance information. Because ocean observatories must tolerate and exploit data and products that weren't created by provenance-aware software, we must allow users to assert facts about data and products.[13] This capability allows the incremental accretion of provenance knowledge as needed—a "pay as you go" approach.[11]
- *Intensional provenance for multigranular data.* There's no single "correct" way to decompose a large heterogeneous data inventory into logical data sets—for example, a year-long timeseries plot carries its own provenance (for example, author, time of creation, or quality control policies applied), but so do the individual measurements included in the plot (such as instrument, calibration settings, and time of delivery). Fine-grained provenance might become too voluminous to manage, but coarse-grained provenance doesn't provide sufficient detail. One solution we're exploring is to track provenance "intensionally." For example, we record that "all measurements made between September and October exhibit biofouling"[14] as opposed to recording an explicit biofouling code with thousands of individual measurements. All applicable provenance assertions might then be included with data and products as they're delivered.

Ocean observatories provide transformative opportunities for understanding and managing one of the world's most complex environments. However, observatories will be only as effective as we can make the scientific exploration of their data and products. The key to effective scientific exploration will be a strategy that ensures that observatory products have known provenance so they can be shared, interpreted, and adapted to new purposes.

The generation of provenance-aware products in ocean observatories pose both common challenges and domain-specific ones. CORIE and SATURN are among selected observatories pioneering the requirements definition and the implementation of solutions in provenance-aware products. These observatories constitute a standing invitation and a challenge for the computer science community to consider theoretical and applied research opportunities in this area.

Acknowledgments

Colleagues and students from different disciplines and affiliations have generously provided the incentive and expertise to explore new ideas and concepts. A core team (Charles Seaton, Paul Turner, Ethan VanMatre, Michael Wilkin, Joseph Zhang, and Bill Howe) operates all facets of the CORIE observatory, under the scientific direction of António Baptista. The US National Science Foundation (OCE-0424602) provided financial support for this research. Research on provenance was partially funded by NSF (OCE-0239072). Research on VisTrails

is also funded by NSF (IIS-0513692, CNS-0541560, OISE-0405402, CNS-0524096, IIS-0534628), the US Department of Energy, and IBM Faculty Awards. Any statements, opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and don't necessarily reflect the views or policies of the federal sponsors, and no official endorsement should be inferred.

References

1. D.L. Martin, "The National Oceanographic Partnership Program, Ocean US, and Real Movement Towards an Integrated and Sustained Ocean Observing System," *Oceanography*, vol. 16, no. 4, 2003, pp. 13–19.
2. H.L. Clark and A. Isern, "The OOI and the IOOS: Can They Be Differentiated? An NSF Perspective," *Oceanography*, vol. 16, no. 4, 2003, pp. 20–21.
3. A. Chawla et al., "Seasonal Variability and Estuary-Shelf Interactions in Circulation Dynamics of a River-Dominated Estuary," to be published in *Estuaries and Coasts*, 2008.
4. D.L. Bottom et al., *Salmon at River's End: The Role of the Estuary in the Decline and Recovery of Columbia River Salmon*, tech. memorandum NMFS-NWFSC 68, National Oceanic and Aeronautic Assoc., 2005.
5. US Army Corps of Engineers, Biological Assessment of Columbia River Channel Improvements Project: An Internal Report to the National Marine Fisheries Service and US Fish and Wildlife Service, US Army Corps of Engineers, Portland District, 2001.
6. Y.L. Zhang, A.M. Baptista, and E.P. Myers, "A Cross-Scale Model for 3D Baroclinic Circulation in Estuary-Plume-Shelf Systems: I. Formulation and Skill Assessment," *Continental Shelf Research*, vol. 24, no. 18 2004, pp. 2187–2214.
7. Y.-L. Zhang and A.M. Baptista, "A Semi-Implicit Eulerian-Lagrangian Finite-Element Model for Cross-Scale Ocean Circulation, with Hybrid Vertical Coordinates," to be published in *Ocean Modeling*, 2008.
8. L. Bright, D. Maier, and B. Howe, "Managing the Forecast Factory," *Proc. 22nd Int'l Conf. Data Eng. Workshop on Workflow and Data Flow for Scientific Applications*, IEEE Press, 2006, pp.64.
9. J. Freire et al., "Managing Rapidly-Evolving Scientific Workflows," *Proc. Int'l Provenance and Annotation Workshop* (IPAW), LNCS 4145, Springer, 2006, pp. 10–18.
10. B. Howe, D. Maier, and L. Bright, "Smoothing the ROI Curve for Scientific Data Management Applications," *Proc. 3rd Biannual Conf. Innovative Data Systems Research*, ACM Press, 2007, pp. 185–195.
11. M.J. Franklin, A.Y. Halevy, and D. Maier, "From Databases to Dataspaces: A New Abstraction for Information Management," *SIGMOD Record*, vol. 34, no. 4, 2005, pp. 27–33.
12. B. Howe, D. Maier, and N. Rayner, "Quarrying Dataspaces: Schemaless Integration of Unfamiliar Information," to appear in *Proc. 24th Int'l Conf. Data Eng., Workshop on Information Integration Methods, Architectures, and Systems*, IEEE Press, 2008.
13. K. Muniswamy-Reddy et al., "Provenance-Aware Storage Systems," *Proc. Usenix Technical Conf.*, Usenix Assoc., 2006, pp. 43–56.
14. C. Archer, A.M. Baptista, and T. Leen, "Fault Detection for Salinity Sensors in the Columbia Estuary," *Water Resources Research*, vol. 39, no. 3, 2003,1060.

António Baptista is the director of the US National Science Foundation (NSF) Science and Technology Center for Coastal Margin Observation and Prediction, and a professor of environmental and biomolecular systems and of computer science and electrical engineering at Oregon Health & Science University. His research interests are in observing, understanding, and predicting coastal-margin processes. Contact him at baptista@stccmop.org.

Bill Howe is a senior research associate at the NSF Science and Technology Center for Coastal Margin Observation and Prediction. His research interests are in the management, integration, visualization, and dissemination of scientific data. Howe has a PhD in computer science from Portland State University. Contact him at howeb@stccmop.org.

Juliana Freire is an assistant professor at the University of Utah. Her research interests include scientific data management, Web information systems, and information integration. Freire has a PhD in computer science from the State University of New York at Stony Brook. She is a member of the ACM and the IEEE. Contact her at juliana@cs.utah.edu

David Maier is the Maseeh Professor of Emerging Technologies in the Computer Science Department at Portland State University, where he also leads the DataLab research group. His research interests include data stream systems, scientific data management, superimposed information, and dataspaces. Contact him at maier@cs.pdx.edu.

Cláudio T. Silva is an associate professor at the University of Utah. His research interests include visualization, geometry processing, graphics, and high-performance computing. Silva has a PhD in computer science from the State University of New York at Stony Brook. He is a member of the IEEE, the ACM, Eurographics, and Sociedade Brasileira de Matematica. Contact him at csilva@cs.utah.edu.