EarthCube White Paper: Advancing Scientific Understanding by Communicating via Data Interactive Publications

Last modified: 17 October 2011

Ben Domenico

Unidata Program Center
University Corporation for Atmospheric Research (UCAR)
Sponsored by NSF Atmospheric and Geospace Sciences (AGS)

(**Note**: This EarthCube whitepaper is an abridged and generalized version of an earlier whitepaper prepared for the Unidata Policy Committee written before the author was aware of the Earthcube initiative. The online version of the paper illustrates the data interactivity aspect of the publication, whereas the Microsoft Word and PDF versions have some live links but the server processing that generates the display figures is not run "on the fly." Printed copies of course are just printed copies. The white paper is also referenced by the OpenGeospatial Consortium as OGC Document 11-146. The reader is encouraged to work with the online version of the whitepaper at https://sites.google.com/site/datainteractivepublications/earthcube-whitepaper-on-data-interactive-publications)

Abstract

Imagine a scientific environment in which authors create online publications that allow readers to access, analyze, and display the data and processes discussed in the publication. Rudimentary examples of such documents can already be cobbled together using existing technological tools in conjunction with the appropriate interface standards. Working together, the science, technology and publishing communities can build on these foundations to develop sophisticated cyber- and organizational infrastructure that will revolutionize how scientists and science educators interact with one another and with the general public. The idea is simple: the reader of a publication will have access not only to the datasets under discussion but also to the processes used by the author to carry out the analysis and display of those datasets. The reader will be able to repeat the experiment as it is published, or perform related experiments by using different datasets or different processes.

Motivation and Vision

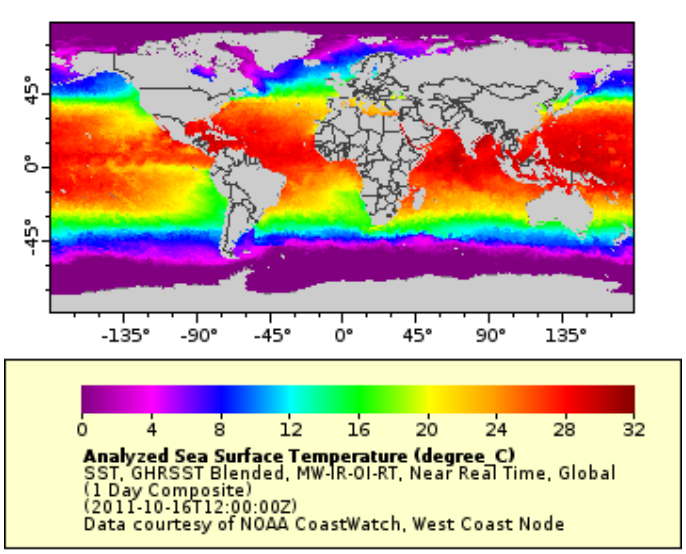
In the academic community, publications are the "coin of the realm" -- arguably the highest-valued metric of accomplishment. Data access and analysis are the basis of the vast majority of these publications. By enabling interaction with the relevant datasets from within the document, data interactive publications will close the gap between the publication process and the research and education processes. Moreover, fully interactive publications which directly cite both the data and the processes involved in the research will improve transparency and reproducibility, which are emerging issues in the geosciences. In this era of large datasets, the ability to house data and processing services on the same system increases scalability by reducing the need to transport data to processing centers. In the education community, many students are complementing desktop applications with browser-based analysis and display tools, many of which can be controlled via mobile platforms such as smartphones, tablets, and e-book readers. As the technological environment moves toward these mobile platforms, computation and data access will move to the server side -- perhaps in the cloud. All of these developments provide a window of opportunity to bridge the gap between scientific publications and the data analysis that goes into them.

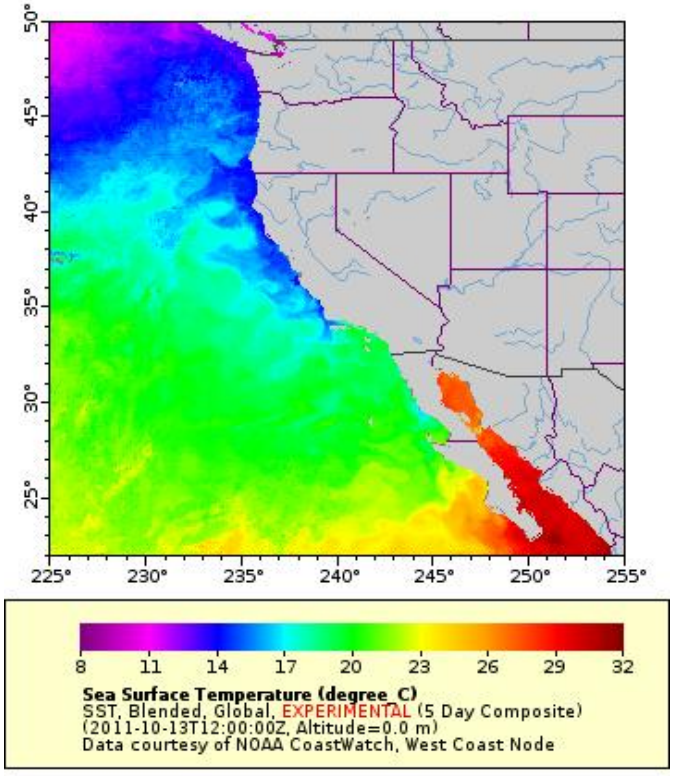
What can be done now

At present, the simple, intuitive, openly available interface of the Google Sites wiki can be used to access data and web data processing services on servers in the community. One such server is at ERDDAP (the Environmental Research Division's Data Access Program at the NOAA Coastwatch site.)

In the ERDDAP examples below, the latest global Sea Surface Temperature (SST) map and a regional experimental and a blended 5-day experimental composite SST are generated on the fly using data and web processing services freely available at the Coastwatch site. As the page loads, the images are automatically generated with the latest data. As noted, the reader can modify the processes that generate the images or download the displayed data -- or related data -- by clicking the appropriate links.

(Be patient. The images below are generated "on the fly" from real data on the ERDDAP server. Depending on server and network activity, it sometimes takes a while to load. A browser page "reload" helps in some cases.)





Clicking on one of the images takes the reader to a page where she can modify the process that generates the image, download the associate datasets or interact with related datasets and perform different calculations. You can also access a table of the data in the map, or download a csv file, a netcdf file , or a matlab file among many other formats.

Currently available (formal and community) standard web services have made it possible to create rudimentary examples of such data interactive documents with the following characteristics:

As the example shows, straightforward HTML documents can be created on a public wiki or local web server. Via embedded links, the reader accesses not only the usual textual information in the publication, but also the data involved in the research, the processes used to analyze the data, and the display tools used to create the illustrations. From within the publication, the reader can interact with data via:

1. Data cited in the publication plus a host of related datasets
2. Analysis tools running on the server or client side
3. “Live” tables with real data in them
4. Downloads of the datasets

Earlier efforts to create such publications required a rich desktop client on the reader's desktop for performing the analysis and display. The advent of web processing services make it possible to access and interact with such documents using a thin web browser client or even a mobile platform "app."

Advantages of Online Data Interactive Publications

Moving science publications into the interactive realm provides a community incentive for creating an additional powerful form of metadata, in the form of the publications themselves. In an era of data interactive documents, the academic is rewarded for "documenting" the data itself. Moreover, this approach fosters transparency and openness by providing broader and easier access to processing and data in addition to textual descriptions and processed figures representing the data. Data discovery is enhanced because the publications themselves can be found via existing web search facilities, which in turn leads the searcher to the datasets and analysis tools embedded in the publications. Search system rankings can be based on pointers into and out of the publications. Flexibility and scalability are enhanced, because services can be located on a local workstation/cluster, on a central server, or in the cloud. Processing can be performed near "big data" stores, reducing data transport requirements.

It is important to realize that, while the entire end-to-end, loosely coupled system of components results in a facility that can transform the fundamental communication mechanism of the academic community, subsets of the overall system can also make fundamental improvements in the way science is conducted in the era of big data. Simply moving computation to servers where the data reside will be a big win in many instances. This transformation will also make it possible to initiate and steer computations involving large data collections from modest, increasingly ubiquitous mobile platforms.

Contributors and Collaborators

Interactions on this topic on the EarthCube web site have made it clear that several organizations are already involved in prototypes and examples of data interactive publication systems. For example, Erik Franklin points to the geosymbio page as one example. Ian Foster indicates he has a graduate student working on this topic. As noted elsewhere, the engagement of the publication industry/community is critical to success in the long run. The Dryad System that Matt Jones cites looks very encouraging in that regard. An important thing to note is that the community is at a juncture where it will be very valuable to gather all the experience, examples, and prototypes that we have and determine whether there are some common elements that communities new to the field might build on. It will be valuable to hear of more success stories (e.g., getting the peer review publications industry involved) where others of us have foundered. And, while there will be tailored systems to serve various science communities and disciplines, there may be a few places where agreeing on specific protocols could be of general benefit, e.g., data citation (which is being addressed elsewhere but has a definite role here as well) and process citation, i.e., referencing an actual web process that was used in an experiment described in a publication. Agreement in some of these area could greatly facilitate cross-disciplinary research.

It should be clear this will not work as a one-size-fits-all mechanism scientific publication -- much less one central system that would work for all science publications. As Walt Snyder points out, it must be tailored to each community.

In addition to the groups interacting on the EarthCube site, others are developing components that can become part of data interactive document systems. Simple examples of data interactive documents using existing web services components -- including the Live Access Server at NOAA's Pacific Marine Environment Labs (PMEL), the ERDDAP system at the NOAA Coastwatch site, the University of Reading WMS/GODIVA combination, and the Unidata Integrated Data Server -- are available now:

1. Main Wiki Site with ERDDAP visualization of sea surface characteristics off the West Coast of the US
2. Original Unidata-centric, pre-EarthCube whitepaper
3. THREDDS / WMS / GODIVA display of temperature forecast from Global Forecast System model
4. Unidata Integrated Data Viewer 3D interaction with the same GFS data
5. Live Access Server sea surface temperature analysis and display
6. Lamont Doherty International Research Institute (IRI) Examples
7. NASA GIOVANNI data display is in the works, but suffering an operator error "glitch" at the moment

Other technologies, such as the Ferret-THREDDS Data Server (PMEL and Unidata) and the GrADS-Data Server out of the Institute for Global Environment and Society (IGES) are also promising.

Remaining Challenges

While the simple examples listed above illustrate the possibilities, there are many challenges to overcome before this approach can become an integral part of the scientific culture. This is a new way of thinking for scientists, publishers, and developers of web services. The difficult issue of ensuring the persistence of online resources is compounded by the added requirements that datasets and processing services are also available in the future. Furthermore, computational resources must be made available in addition to data access; this in turn requires secure but convenient authentication and authorization services. More sophisticated web processing services will be needed for publication based on complex algorithms. Many of today's web processing services are too new and untried for people to rely on them. Finally, it is crucial that the scientific publication industry be engaged in the process.

As might be expected, one of the areas most in need of creative ideas is how to fund the resources needed to operate in a world where the bulk of the computation needed for analysis and display moves from desktop personal computers to the cloud. Industry giants such as Google, Microsoft, and Amazon have implemented different approaches for recovering costs. These range from Google advertising, to Microsofts addition of cloud-based augmentation (Office 365) of traditional desktop "productivity" programs. A crucial question for the academic community is how to pay for and allocate resources that are accessed in the cloud.

Building Blocks for The Foundation

Many of the pieces are already in place to begin working toward the long term goal of data interactive scientific publications. In particular, the standards of the OpenGeospatial Consortium (OGC) make it possible to serve and access datasets via agreed upon web services protocols such as the Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), and Sensor Observation Service (SOS). They have also defined standard encoding formats such as the Geography Markup Language (GML) with specialized dialects applicable in the geosciences, e.g., the Climate Sciences Modelling Language (CSML)and the emerging WaterML. For binary encoding, the network Common Data Form (netCDF) has recently been adopted as an OGC standard. So there exist standard forms for the payloads delivered via the standard protocols. In addition brokering approaches can now be employed for combining the various discovery, data access and web processing protocols into sophisticated systems of systems. Recent OGC emphasis on Web Processing Services (WPS) and REST (REpresentational State Transfer) forms for the interfaces specifications are important additional foundational steps, because they enable access to computational resources via URL-based references that can be embedded into online publications.

Possible Next Steps

Building on the foundation of standard interfaces can take place within the existing geosciences communities. CUAHSI (Consortium of Universities for the Advancement of Hydrological Sciences Inc.) is a leading player in the development of WaterML. The atmospheric and ocean sciences have a long history of services and applications based on netCDF. Within research and education groups like the Unidata community of university departments, one can envision a phased augmentation of desktop analysis and display tools like the Integrated Data Viewer (IDV) running in a computing lab of individual workstations to an environment where data analysis processing functions are implemented on a departmental cluster but are initiated and controlled by students and faculty via ubiquitous mobile tablets, smartphones and e-book readers. This incremental approach could provide hands-on experience with the technology of interactive remote computation while the academic community seeks solutions to the long term cultural and financial challenges of distributed interactive publication systems. Getting there will require a concerted and coordinated effort of many organizations: funding agencies, university libraries, members of the publications industry, the geosciences research and education community, as well as the developers of required cyberinfrastructure.

Related Reading

AMS Data Interactive Extended Abstract, Jeff Weber, Ben Domenico
http://ams.confex.com/ams/pdfpapers/87619.pdf

Brokering Approaches to Earth Science Cyberinfrastructure
http://earthcube.ning.com/group/technology-solutions/forum/topics/the-brokering-approach-for-earth-science-cyberinfrastructure

Climate Science Modelling Language (CSML)
http://ndg.badc.rl.ac.uk/csml/

Consortium of Universities for the Advancement of Hydrological Science (CUAHSI)
http://www.cuahsi.org/

DRYAD, suggested by Matt Jones
http://datadryad.org/

Early Examples (some of which no longer work for reasons cited in this whitepaper)
http://www.unidata.ucar.edu/projects/THREDDS/DataPublications/

ERDDAP, NOAA Coastwatch, examples facilitated by Roy Mendelssohn
http://coastwatch.pfeg.noaa.gov/erddap/index.html

Ferret-THREDDS Data Server (F-TDS)
http://ferret.pmel.noaa.gov/LAS/documentation/the-ferret-thredds-data-server-f-tds/

GeoSymbio, suggested by Erik Franklin
https://sites.google.com/site/geosymbio/

Geography Markup Language (GML)
http://www.opengeospatial.org/standards/gml

Integrated Data Viewer, Unidata
http://www.unidata.ucar.edu/software/idv/

Live Access Server, NOAA PMEL
http://ferret.pmel.noaa.gov/LAS

NetCDF, OGC Specification
http://www.opengeospatial.org/standards/netcdf

NetCDF, Unidata Documentation
http://www.unidata.ucar.edu/software/netcdf/

ncWMS, U of Reading, written by Jon Blower
http://www.resc.rdg.ac.uk/trac/ncWMS/

OpenGeospatial Consortium (OGC)
http://www.opengeospatial.org/

Original White Paper from Unidata Policy Committee Presentation
https://sites.google.com/site/datainteractivepublications/home/white-paper-on-data-interactive-publications

REpresentational State Transfer (REST)
http://en.wikipedia.org/wiki/Representational_state_transfer

Sensor Observation Service (SOS)
http://www.opengeospatial.org/standards/sos

THREDDS Data Server, Unidata
http://www.unidata.ucar.edu/software/tds/

Unidata Program Center
http://www.unidata.ucar.edu/

WaterML
http://www.opengeospatial.org/projects/groups/waterml2.0swg

Web Coverage Service (WCS)
http://www.opengeospatial.org/standards/wcs

Web Feature Service (WFS)
http://www.opengeospatial.org/standards/wfs

Web Map Service (WMS)
http://www.opengeospatial.org/standards/wms

Web Processing Service (WPS)
http://www.opengeospatial.org/standards/wps