International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Data Provisioning for the Object Modeling System (OMS)

[1] Jack R. Carlson, [12] Olaf David, [12] Wes J. Lloyd, [1] George H. Leavesley, [4] Ken W. Rojas, [3] Timothy R. Green, [1] Mazdak Arabi, [1] Lucas Yaege, and [1] Holm Kipka

[1] Colorado State University, Dept. of Civil and Environmental Engineering
Fort Collins, Colorado 80523 USA
Jack.Carlson@colostate.edu
[2] Colorado State University, Dept. of Computer Science
Fort Collins, Colorado 80523 USA
[3] USDA-ARS-NPA, Agricultural Systems Research Unit
2150 Centre Ave., Bldg. D, Suite 200, Fort Collins, Colorado 80526 USA
[4] USDA-NRCS Information Technology Center
2150 Centre Ave., Building A, Fort Collins, Colorado 80526 USA

Abstract: The Object Modelling System (OMS) platform supports initiatives to build or re-factor agro-environmental models and deploy them in different business contexts as model services on cloud computing platforms. Whether traditional desktop, client-server, or emerging cloud deployments, success especially at the enterprise level relies on stable and efficient data provisioning to the models. In this paper we describe recent experience and trends with tools and services to supply data for model inputs. Solutions range from simple pre-processing tools to data services deployed to cloud platforms. Also, systematic, sustained data stewardship and alignment with standards organizations impart stability to data provisioning efforts.

***Keywords**:* data provisioning; environmental modelling; data services, model services, cloud computing

1 INTRODUCTION

The Object Modeling System (OMS) described by David et al (2013) provides a framework for building or re-factoring models to assist analysis and decision-making for land management supported by conservation programs. OMS also provides a platform for deploying these models in different enterprise-level business contexts. For example, Leavesley et al (2010) discuss OMS data and model services developed to support twice-monthly water supply forecasting on 600 basins in the western United States. Lloyd et al (2012b) describe erosion model services for daily conservation planning across 2,800 county offices in the country, Leavesley et al (2014) discuss OMS-based monthly water balance modelling across nine countries of the Nile Basin, extensible to other regions of the world.

A primary constraint on the efficient, stable, and timely use of these models in enterprise deployments involves provisioning data from disparate data sources. Data flow involves a three-way interaction between data services, model services, and the business application integrating the services. We define service to include both the data and the processes that act on it. We define enterprise to mean a large public, private, or hybrid sector organization having moderate to heavy daily computing demand and business to mean the processes the organization applies to carry out its mission.

This paper examines the process of provisioning input data for enterprise-level OMS-based model applications involving climate, water, soil, vegetation, and land management data at farm/field, small watershed, and basin scales.

2 ANALYSIS OF THE DATA PROVISIONING CONSTRAINT

To help organize its data provisioning strategy, the OMS team has analysed current resources and practice across organizations in the agro-environmental domain, highlighted in this section.

2.1 Data Services

Among many sources for model input, important data stores for OMS model services include:

Applied Climate Information System (ACIS, http://rcc-acis.unl.edu)
Climate Research Unit (CRU, http://www.cru.uea.ac.uk)
Ecological Site Information System (ESIS, https://esis.sc.egov.usda.gov0
Gridded Vegetation Indices (MODIS NDVI/EVI, http://modis.gsfc.nasa.gov)
Land Management Operations Database (LMOD)
National Water Information System (NWIS, http://waterdata.usgs.gov/nwis)
PLANTS (http://plants.usda.gov)
Snow Telemetry (SNOTEL, http://www.wcc.nrcs.usda.gov/snow)
Soil Survey Geographic Database (SSURGO, http://datagateway.nrcs.usda.gov),
Water Quality Exchange (WQX/STORET, http://www.epa.gov/storet)

Access to this data has trended from manual distribution on electronic media to online dataset downloads to Simple Object Access Protocol (SOAP) data access web services and then to more contemporary Representative State Transfer (RESTful) services. The most efficient providers enable very specific data requests specified in a Hypertext Transfer Protocol (HTTP) GET request matching the input requirements of the model service, and provide tools for building the request. For example, see NWIS at http://waterservices.usgs.gov/rest/Site-Service.html. Less streamlined access increases the burden on the business application or model service to process and fit the data to the model's requirements.

Although not apparently a current problem, data service availability and scalability becomes more important as models are integrated with enterprise-level business applications with large user bases. Data providers should be able to scale-out their services and provide fail-over in order to provide expected quality of service (QoS) through a service level agreement (SLA) process. If important enough for QoS and permitted, an enterprise may mart an instance of the data service internally for their use.

An enterprise should factor in expected data service longevity. Do data dictionaries align with standard vocabularies? Is the data service supported by a well-organized stewardship organization and process? Does the data provider have a good track record for service?

2.2 Model Services

An older legacy model can be deployed as a black box executable within a model web service. Deploying more than one model in this manner with a common data service for input usually requires data translation code for each model. For example from Muth and Bryden (2013), a legacy water erosion model and legacy wind erosion model may run against the same data service for soil and land management inputs. Each model consumes the data differently and therefore it must be translated to the model's requirements. To the extent possible new model development should try to avoid the need for translation.

Enterprise deployments of multiple models in a business application should have a consistent way to consume data across model services. Model services also should employ techniques to minimize round-trips to get data.

2.3 Business Applications

In examining data provisioning approaches described by Winchell et al (2007), Johnston et al (2011), Ames et al (2012), Rosenzweig et al (2013), and Werner et al (2013), we find most data pre-processing for model input occurs in the business application using a model service, whether in the

application itself or a plug-in or companion application. Many applications contain geospatial components and tools to transform and associate data with response units on a map (e.g. hydrologic response units, or HRUs). These applications usually provide a tool for creating and editing response unit geometry and attributes, and base layers for response unit delineation and backdrop.

In our experience, business applications also mediate requests to data services, sometimes returning choices for the user to choose for model input, for example, the soil component and its attributes for the model run. Applications usually enable the user to edit certain input elements in the model parameter file, such as replacing default soil component slope length and steepness values.

Persisting pre-processing code and components in an open repository obviously encourages re-use. Where possible general-purpose scripts and tools should be designed to process different kinds of similarly formatted data, for example different kinds of gridded data.

3 STEPS TO EFFECTIVELY PROVISION OMS MODEL SERVICES

Effective data provisioning involves continuous improvement and commitment throughout modelling enterprises and data providers. From the as-is analysis, the OMS team has organized its data provisioning strategy around the following steps.

3.1 Standardize Data Provisioning Architecture

Legacy models re-factored or wrapped as OMS model services sometimes come with their own data stores. For example, the desktop version of the Revised Universal Soil Loss Equation (RUSLE2) consumes land management, soil, and climate data contained in old .gdb formatted files. The desktop version of the Wind Erosion Prediction System (WEPS) consumes some of the same data contained in files of other formats. The Soil Condition Index (SCI) calculation usually involves both RUSLE2 and WEPS runs. Converting RUSLE2 and WEPS to web services, and combining both models in a SCI web service requires a more efficient data provisioning architecture for model input to reduce duplication and data management support.

Figure 1 shows the basic construct of the OMS data provisioning architecture. A stewardship group keeps soil, climate, or other data current in a warehouse, which feeds one or more data marts designed to support OMS model services. A business application connects to relevant data services and the model service to mediate the flow of input data to the model.

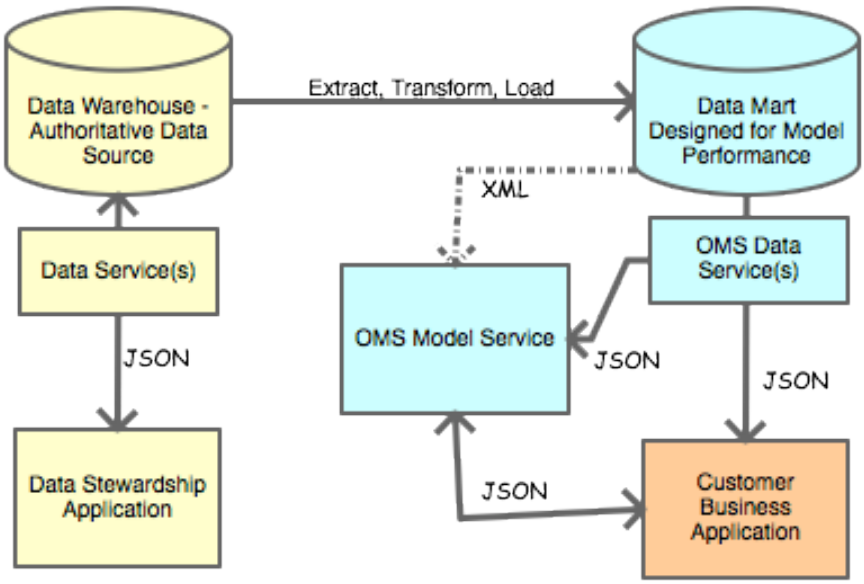

Figure 1. Conceptual data provisioning architecture for OMS model services

Ideally, a source provider manages authoritative data in a warehouse data store using a stewardship application with appropriate approval authorities. The provider designs the warehouse to facilitate stewardship and employs extract, transform, and load (ETL) tools to feed a data mart designed to efficiently provision OMS model services. Sometimes a business application requests a data payload from the data mart before editing and sending to the model service. The model service may receive an input payload from the data mart's web service, or the model may get data directly from the mart, for example a URL link to an Extensible Markup Language (XML) file in a data store on a web server. Data can be stored in many different formats and structures, but the basic flow of data for provisioning model services in Figure 1 applies. Some OMS-related projects involve managing the entire data provisioning workflow, and some only partially so. The latter case assumes the same general architecture, requiring service level agreements or some reasonable assurance of the quality and reliability of the data source and services.

3.2 Integrate with GIS Applications to Facilitate Creation of Response Units

Many if not most agro-environmental models operate across a series of landscape units. Modelers often refer to them as response units, areas of land to which a model associates its output. Therefore a response unit also contains a unique set of model input data and provides a crucial organizing entity for processing source data for model input. Effective applications running these model and data services must contain geospatial (GIS) processing components. OMS currently integrates with three geospatial platforms containing tools for response unit delineation and pre-processing data for response unit-based model input.

The Environmental Resource Analysis and Management System (eRAMS) integrated with OMS described by Wible and Arabi (2013) provides a geospatial application development and data management platform for scalable model services. Using eRAMS, Leavesley et al (2014) have developed an automated process to create basins, sub-basins, and hydrologic response units (HRUs) from a digital elevation map (DEM) and monitoring stations, followed by automated processes to generate HRU parameters and input data for an OMS water balance model service. The JGrass-NewAge hydrologic modeling system integrates OMS with the JGrass-based uDIG GIS and visualization platform, discussed by Formetta et al (2014). The Geospatial Modeling Interface (GMI) from Ascough II et al (2012) contains geospatial tools for OMS model simulation set-up and visualization.

3.3 Develop and Maintain Pre-Processing Tools to Generate Input Data

OMS provides a simple model service input convention using the comma separated value (CSV) standard for two types of data: table and property, which are annotated by @T and @P respectively. For tables, @H annotates header information, and for properties @S annotates sections containing properties. Both tables and properties can be included in the same data file.

# Table example **@T, "Example DataSet"** **CreatedAt, 5/11/12** **CreatedBy, Gary Nelson** **# Now, there is header information** **@H, time,b,c** **Type, Date,Real,Real** **Format, yyyy-MM-dd,#0000.00,#000.0000** **,2006-05-12,0000.00,001.1000** **,2006-05-13,0001.00,002.1000**	**# Property example** **@S, "Parameter"** **CreatedAt, "Jan 02, 2013"** **CreatedBy, Joe Smith** **# Single Properties** **@P, coeff, 1.0** **description, "A coefficient"** **public** **@P, start, "02-10-1977"** **description, "start of simulation"**

Table 1. OMS table and property convention for model input data.

For water supply forecasting in the western U.S., the OMS team recently has employed eRAMS to access ACIS and SNOTEL data services to retrieve meteorological data, NWIS data services for streamflow data, and then applied rapidly developed Groovy scripts (http://groovy.codehaus.org) to update and slice the data into OMS-compliant .csv flles. Other Groovy scripts process this data and distribute to the HRUs in the selected forecast basin, using either the XYZ distribution method by Hay et al (2000) or de-trended kriging method by Garen et al (1994).

Leavesley et al (2014) discuss a script developed using Python (https://www.python.org) adapted to extract monthly climate data from CRU to create OMS-compliant .csv files for precipitation, maximum and minimum temperatures, potential evapotranspiration, and relative humidity for water balance modeling in the Nile Basin. Another script run on other data sources creates .csv files of monthly values for H2 and O18 isotope concentrations, NDVI, soil water holding capacity, and vegetative cover density.

Streamlined pre-processing relies in part on adapting and re-using these and other scripts and tools created from application to application. Therefore they have been posted to the OMS Component Library at http://omslib.javaforge.com.

3.4 Deploy Data Services to a Cloud Platform

OMS managed data services, as well as model services, operate in the OMS Cloud Services Innovation Platform (CSIP), described by Lloyd et al (2012a).

SSURGO soil data services run in CSIP against a 320 gigabyte SSURGO PostgreSQL/Postgis database. The database is horizontally partitioned into shards on 8 virtual machines (VMs). Stress testing of the data service for 1-2 thousand concurrent user sessions from a 10,000-user community has projected a requirement for 2 VMs. The Java API for RESTful Services (JAX-RS) soil data service intersects application provided location with soil mapunit geometry and returns requested parameter data in a JavaScript Object Notation (JSON) payload. The current architecture supports less than 10 millisecond query response.

David et al (2014) describe the 2 gigabyte Land Management Operations Database (LMOD) deployed to CSIP in a PostgreSQL database. JAX-RS RESTful JSON-based data services include returning a list of managements and returning parameters for a selected management.

Kipka et al (2013) have developed a data service now called LAMPS supporting the creation of an LMOD-based land management system from annual cropping imagery in the USDA-National Agricultural Statistics Service (NASS) Cropscape system. Among other uses, the service will be applied for analysing resource concerns for benchmark conditions.

To date OMS model services require relatively modest data inputs, involving uncomplicated data queries. Should complexity increase, performance modelling by Lloyd et al (2012) provides insight for optimizing the data service architecture to eliminate bottlenecks. To this point deployed OMS model services also have not encountered significant lack of availability or delay getting data for model input from external providers.

OMS does not currently use NoSQL data store technologies for persisting model input data, but leverages Memcached (http://memcached.org) and Redis (http://redis.io) key/value stores to cache requests/response objects during OMS model service runs. In some cases, static model input data has been stored as XML files on web servers (e.g. Apache http://httpd.apache.org, nginx, http://nginx.org) for quick retrieval using cURL (http://curl.haxx.se) scripts.

3.5 Ensure Commitment to Data Stewardship

SSURGO has been an integral part of the USDA Natural Resources Conservation Service (NRCS) Soil Survey Program for many years. NRCS manages soil data in a very systematic manner through its network of soil survey offices and soil scientists and an integrated system for acquiring, processing, warehousing, and distributing data for internal and public use. The OMS-CSIP SSURGO database comes from this system, the data source for the model services used internally by the agency.

NRCS also performs data stewardship for LMOD through its network of national, regional, and state agronomists. The agency will deploy its new online Stewardship Management Application (SMA) for LMOD in 2014, replacing the legacy desktop application approach.
Enterprise application systems integrating external data services, especially with large user bases and critical business processes, may require formal service level agreements (SLAs) containing

commitment to keep the data current on an agreed upon schedule. Otherwise, an organization assumes the risk data becomes stale over time and no longer usable.

3.6 Align Data Stores and Services with Standards Organizations

OMS attempts to associate with formal or de facto standard data stores to stabilize the data provisioning process. National Cooperative Soil Survey standards underpin the OMS/CSIP SSURGO data mart. LMOD agronomic-oriented data definitions are being integrated with the AgGateway Field Operations initiative (www.aggateway.org), bridging with International Standards Organization (ISO) machine-oriented data entities. Integration will enable data exchange across the agricultural domain: farmer, consultant, agri-business, and government agency.

4 OMS-MANAGED DATA STORES

For long-term availability and performance, the OMS team manages a core set of natural resource related data marts for provisioning model inputs to currently deployed and planned CSIP model services.

4.1 Land Management Operations

LMOD contains 55,736 land managements, 3,279 crops (vegetations), 1,082 operations, 99 wind barrier practices, 39 contouring practices, 403 strip/barrier practices, 195 residues, and 30 fuels. LMOD groups land managements (e.g. cropping systems) into 75 crop management zones (CMZs) across the U.S. LMOD contains 639 parameters, 393 in use for four model services: sheet/rill erosion (RUSLE2), wind erosion and air quality particulate matter (WEPS), soil condition index (SCI), and soil tillage intensity rating (STIR). Going forward LMOD will support model services for pesticide hazard, nutrient balance, runoff and groundwater management, irrigation scheduling, and possibly grazing schedules. The essential structure of LMOD involves land managements having a schedule of crops/vegetation and operations, and practices impacting the landscape.

4.2 Soil

The CSIP deployment of SSURGO contains ~30 million soil mapunit polygons and their soil survey attributes. The SSURGO data services will be extended to support the new model services described for LMOD above.

4.3 Climate

Currently, LMOD also stores 10,710 climate records containing data inputs for the RUSLE2 model service. These records will be separated from LMOD into a separate data store with web services, and likely integrated with other climate data stores managed by the NRCS Water and Climate Center. The WEPS model contains climate (CLIGEN) and wind (WINDGEN) generators, which will be separated and deployed as separate data services.

4.4 Other

The OMS team has been tasked to design, build, and deploy on-line data marts for nutrients, pesticides, livestock, wildlife, and ecological sites to support conservation planning and application tools.

5 SUMMARY AND CONCLUSIONS

Business applications running OMS model services get input data from a variety of sources, from data stores in OMS-CSIP and externally from NWIS, ACIS, and elsewhere. Enterprise deployments favour data services that are stable, highly available, accessible, and performant. Stability usually reflects long-term commitment and support to data stewardship by the provider and alignment with standards organizations. A common availability metric for SLAs specifies “three-nines”, 99.9% up time.

Accessibility reflects the ability of the data service to match the input requirements of the model. And performance usually means query response times in milliseconds for local stores and seconds for external stores.

For agricultural and environmental models the most time consuming data provisioning process often involves pre-processing and distributing data across map-based response units. The affected business applications must integrate a sufficiently featured GIS and if possible, broad spectrum pre-processing tools to fully automate and streamline this process. An area that continues to have business value, but likely to trend to exchange of open source scripts and tools.

Finally, data access will continue to trend towards lightweight REST-based services as models are increasingly deployed as services to cloud infrastructures.

REFERENCES

Ames, D., Horsburgh, J., Cao, Y., Kadlec, J., Whiteaker, T. Valentine, D., 2012. HyrdoDesktop; web-services based software for hydrologic data discovery, download, visualization, and analysis. Environ. Modell. Softw. 37,146-156.

Ascough II, J., David, O., Murthy, S., 2012. The AgESGUI geospatial simulation system for environmental modeling application and evaluation. IEMSS 2012 International Congress on Environmental Modelling and Software. Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany, pp. 1519-1526

David, O., Ascough II, J., Lloyd, W., Green, T., Rojas, K., Leavesley, G., Ahuja, L., 2013. A software engineering perspective on environmental modeling framework design: the Object Modeling System. Environ. Modell. Softw. 39, 201-213.

David, O., Rojas, K., Yaege, L., Lloyd, W., Carlson, J., Ascough, J., Green, T., Geter, F., 2014. The Land Management and Operations Database (LMOD). IEMSS 2014 International Congress on Environmental Modelling and Software, Bold Visions of Environmental Modelling, Seventh Biennial Meeting, San Diego, California, USA. 8 p.

Formettta, G., Antonello, A., Franceschi, S., David, O., Rigon, R., 2014. Hydrological modelling with components: a GIS-based open source framework. Environ. Modell. Softw. 55, 190-200.

Garen, D., Johnson, G., Hanson, C., 1994. Mean areal precipitation for daily hydrologic modeling in mountainous regions. Water Resources Bulletin 30(3), 481-491.

Hay, L., Wilby, R. Leavesley, G., 2000. A comparison of delta change and downscaled GCM scenarios for three mountainous basins in the United States Journal of American Water Resources 36(2), 387-397.

Johnston, J., McGarvey, D., Barber, M., Laniak, G., Babendreirer, J., Parmar, R., Wolfe, K., Kraemer, S., Cyterski, M., Knightes, C., Rashleigh, B., Suarez, L., Ambrose, R., 2011. An integrated modeling framework for performing environmental assessments: Application to ecosystem services in the Albemarle-Pamlico basins (NC and VA, USA). Ecological Modeling 222, 2471-2484.

Kipka, H., David, O., Lyon, J., Garcia, L., Green, T., Ascough II, J., Rojas, K., 2013. A web-service tool to generate crop rotation management input files for spatially distributed agroecosystem models, Colorado State University Hydrology Days 2013, 38-46.

Leavesley, G., David, O., Garen, D., Goodbody, A., Lea, J., Marron, J., Perkins, T., Strobel, M., Tama, R., 2010. A modelling framework for improved agricultural water-supply forecasting. 2nd Joint Federal Interagency Conference, Las Vegas, Nevada, June 27-July 1, 2010.

Leavesley, G., Belachew, D., David, O., Patterson, D., Carlson, J., Aggarwal, P., Arabi, M., 2014. Deployment of a water balance model with isotopes (IWBMIso) using eRAMS. IEMSS 2014 International Congress on Environmental Modelling and Software, Bold Visions of Environmental Modelling, Seventh Biennial Meeting, San Diego, California, USA. 8 p.

Lloyd, W., David, O., Lyon, J., Rojas, K., Ascough II, J., Green, T., Carlson, J., 2012a. The Cloud Services Innovation Platform - enabling service-based environmental modelling using infrastructure-as-a-service cloud computing. IEMSS 2012 International Congress on Environmental Modelling and Software. Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany, pp. 1208-1215.

Lloyd, W., Pallickara, S., David, O., Lyon, J., Arabi, M., and Rojas, K., 2012b. Performance modeling to support multi-tier application deployment to infrastructure-as-a-service clouds. IEEE Fifth International Conference on Utility and Cloud Computing (UCC), pp. 73-80.

Muth, D., Bryden, K., 2013. An integrated model for assessment of sustainable agricultural residue for bioenergy systems. Environ. Modell, Softw. 29, 50-69.

Rosenzweig, C., Jones, J., Hatfield, J., Ruane, A., Boote, K., Thorburn, P., Antle, J., Nelson, G., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D. Baigorria, G., Winter, J., 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agric. and Forest Meteorology 170, 166-182.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., Heynert, K., 2013. The Delft-FEWS flow forecasting system. Environ. Modell. Softw. 40, 65-77.

Wible, T., Arabi, M., 2013. Comprehensive flow analysis using cloud-based cyberinfrastructure, Colorado Water 30(5), 15-17.

Winchell, M., Srinivasan, R., Di Luzio, M., Arnold, J., 2007. ArcSWAT Interface for SWAT2005 User's Guide. Blackland Research Center, Texas Agricultural Experiment Station, and Grassland, Soil, and Water Research Laboratory, USDA Agricultural Research Service, 436 p.