International Congress on Environmental Modelling and Software

4th International Congress on Environmental Modelling and Software - Barcelona, Catalonia, Spain - July 2008

Jul 1st, 12:00 AM

Community modelling, and data-model interoperability

Alexey Voinov

Ilya Zaslavskiy

David Arctur

C. J. Duffy

Ralf Seppelt

iEMSs 2008: International Congress on Environmental and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental and Software Society (iEMSs), 2008

Community modelling, and data-model interoperability

Alexey Voinov[a], Ilya Zaslavskiy[b], David Arctur[c], Chris Duffy[d] and Ralf Seppelt[e]

a – Chesapeake Research Consortium Community Program, Johns Hopkins University, USA,
alexey.voinov@uvm.edu
b – San Diego Supercomputer Centre, CA, USA
c – OGC Interoperability Institute, Austin, Texas, USA
d – Penn State University, PA, USA
e – Centre for Environmental Research Leipzig-Halle, Germany

Abstract: Community modelling is a promising paradigm to develop complex evolving and adaptable modelling systems that can share methods, data and models more easily within specialized communities. Why then are cooperative modelling communities still quite rare and do not propagate easily? Why has open source been so successful for software development, yet open models are still quite exotic? One difference between software and models is that software shares some common language. Models often use very different principles, theories, and semantics. For example hydrodynamic models, ecological models, and decision support models may have limited commonalities, In these cases, the disciplinary problem being solved may be the impediment to communication and to development of effective community tools these principles to another; it becomes difficult for one model to talk to another one. Similar problems prevail in data operations, when data sets (which are also models of sort) are hard to integrate with other data.

An issue of contemporary interest is how will community data and models be implemented within environmental observatories. The environmental observatory may are become the ultimate driver for advancing research with a clear need for interoperability standards and functionality.

There are at least three facets to the problem:

- Lack of common modelling and software tools to enable modularity and connectivity;
- Insufficient community understanding or access to basic tools;
- Lack of social motivation and communication skills to enable communal work and sharing environments.

The goals of this paper are to explore these areas with respect to the following points:

- Understand the interoperability needs of the community for data and models within a participatory and collaborative framework;
- Discuss research scenarios that would benefit from interoperability and explore interoperability architecture and standards supporting these scenarios;
- Explore environmental system observatory ontologies, with particular attention to mapping variables to concepts;

- Discuss common access protocols, enabling models to automatically search for data needed and link to data servers. Design data interoperability for model input/output to help link models.

Keywords: Modularity, observatory, culture of sharing, ontology, open source, environmental management

1. INTRODUCTION

There is an increasing consensus that value can be derived from integrating different models and data sets for predictive understanding of environmental problems and for operational forecasting.

Linking models and coordinating data sets has become a leading driver for research at environmental observatories. The nation's ability to monitor, understand and forecast environmental change are a part of NSF strategy for building capacity for research communities engaged in advanced continuous monitoring and modelling of the environment, e.g. CLEANER [1], CUAHSI [2], NEON [3], GEON [4, 28], LTER [5]. They are envisioned as collaborative interoperable systems for synthesizing environmental data from data repositories and real time monitoring into a shared data integration and analysis environment that supports multidimensional and multifaceted modelling of environmental processes, linkage or coupling to environmental sensors/data, and accessible to a broad community of environmental researchers. However they have not been doing much in terms of model integration.

There is also considerable interest in making models "talk" to each other. OpenMI is one such approach.

Other approaches allow data and environmental models to be loosely coupled, tightly coupled, or embedded. Bhatt et al. (this volume) describe how a hydrodynamic model for watershed hydrology greatly improves the modelling process when tight coupling between the GIS tool and the physical model is implemented. Embedded approaches will allow the user to "steer" the physical model in real time such as for development of climate scenarios for long term water resources decision making. Traditional approaches where the GIS tool is independent of the physical model application will continue to be important for simple models with limited data demands, but coupled approaches are likely to be of greater interest researchers and managers.

Development of an integrated data sharing infrastructure to facilitate multidisciplinary collaborative analysis and modelling in the context of an environmental observatory is a pressing need. With such infrastructure, researchers should be able to publish and document their data, discover what information is available based on agreed-upon metadata descriptions, retrieve the information over common data access mechanisms, understand and resolve semantic discrepancies between datasets, integrate them for use in analysis and modelling codes, and share research findings with community members. In this research cycle, information sharing and re-use are the major underpinnings in reducing fragmentation in environmental research, and engaging a broader research community from the environmental science and related domains in advanced data collection, analysis and modelling.

The need to focus on the common data foundation for the communities involved in environmental monitoring, analysis and modelling, is underscored by the following observations:

- Research groups and communities are not isolated but have "neighbours" that are interested or in need of the same data sets;
- Processes and problems are not isolated but have an inherent complexity that requires multi-disciplinary teams need to understand or solve;

- Increasing size and complexity, and hence information needs, of environmental models mean pressure to find automated ways to interpret larger and more complex sets of data across disciplines;
- Besides re-using data collected in neighbouring research domains, it becomes common to include longer time series data in environmental models, which requires that data are archived in such formats and preservation environments that support long-term storage and retrieval (i.e., "data interoperability in time", in addition to interoperability across domains);
- Interoperability focus is a way to consolidate the entire community around common data interchange issues, including data interchange standards and standards governance, heterogeneities across datasets and techniques for resolving such heterogeneities, leading to formulation of new research questions that could not be addressed over a fragmented infrastructure.

The range of interoperability challenges, derived from differences in structure and semantics of datasets, data publication, discovery and access mechanisms, as well as in modelling approaches, have been described in recent literature on environmental observing systems [60]. Technical interoperability issues, such as those related to common procedures for real time data management, integration of streaming data with data archives, and technologies for expressing and resolving well-understood structural and semantic heterogeneities, have been the focus of NSF attention over the last several years, within such initiatives as CEO:P (Cyberinfrastructure for Environmental Observatories: Prototypes), Geoinformatics, CLEANER/WATERS [1, 6] and NEON [3]. However, purely technical solutions for interoperability are insufficient for establishing a shared interoperable infrastructure for environmental observatories. For the software infrastructure to be truly useful for empowering the entire research community with data sharing and collaborative research capabilities, several additional challenges must be addressed. They include:

- making the community aware of the available data and software resources, common component software, and advances within other disciplines (e.g. high performance computing, visualization, etc.)
- building consensus about information models used in the community,
- understanding and harmonizing data structures and data access mechanisms and formats used within the community, and
- building support for modelling applications that take advantage of the infrastructure.

The initial experience of infrastructure development and adoption within GEON [4] and CUAHSI [2] HIS (Hydrologic Information System) projects provides ample evidence of heterogeneities in data and resources needed to compute watershed and estuary models in particular, and of the range of interoperability challenges stemming from different data structures and semantics adopted within the community. Similar interoperability issues are being addressed, in a cross-country setting, within the European "Water Information System for Europe" (WISE) project [27]. At the global scale, the issues of standard formats and exchange protocols for observations data, environmental observations in particular are being addressed by the Open Geospatial Consortium [38,59, 61, 62-65].

Development of ecosystem models in general has been limited by the ability of any single team of researchers to deal with the conceptual complexity of formulating, building, calibrating, and debugging complex models. The need for collaborative model building has been recognized in the environmental sciences. The current-generation models tend to be "idiosyncratic monoliths that are comprehensible only to the builders" [26]. Communicating the structure of the model to others can become an insurmountable obstacle to collaboration and acceptance of the model. The interoperability functions that we propose to develop will be the core of a system of middleware that would allow integrating existing models and will provide for easy integration of new models.

An additional reason for the current disconnect is the mismatch between the generality and scalability expectations of modeling applications, which typically aim at providing the most accurate description of physical processes to support understanding, forecasting and/or decision-making for a fairly narrow well-defined case, vs. development of common software tools, where attaining a generic treatment applicable to a large class of use cases is a necessary ingredient.

Moreover there is a clear need to learn to communicate not just the models, but also be able to dig inside the model structure and have the functionality to extract the most successful pieces and modules to use them in other models. Otherwise we constantly run into the “reinventing the wheel” syndrome, when almost similar functions and algorithms are use to describe ecological processes simply embedded in different data and user interfaces. We could imagine much more advances in modelling if instead of rebuilding and reprogramming the same modules over again in new models, we could focus on further exploration of new ecological phenomena and modelling methods simply plugging in the already tested and approved modules from already existing models.

The models provide ample evidence for heterogeneities that need to be resolved in order to obtain more detailed and accurate model results. There is already a significant community building effort going on. For example, within the Chesapeake Bay area there is the CCMP (Chesapeake Community Modelling Project), which is building reliable working relationships among all participating modellers who are willing to share their model and data structures and needs to develop the important interoperability functions. This will include the integration of the very large existing and continuously increasing watershed, hydrodynamic, and biogeochemical (water quality) databases. This effort is led by the Chesapeake Research Consortium (CRC), which organizes universities, as well as government and non-profit research organizations around common problems. In addition, the CCMP can serve as the body responsible for governance of data exchange standards within the Bay area, providing linkages between various community strata grouped according to:

- organizations: government, industry, research universities and students, environmental and other non-profit groups (including standards groups), citizens, schoolchildren, etc.;
- domains: atmospheric, surface water, soils/geology/landscape, ground water, oceanic, socio-economics, demographics, etc.;
- paradigms: from data collectors/archivists/publishers to modellers, analysts, and decision makers.

Other examples are the NSF-funded CSDMS (Community Surface Dynamic Modelling System - http://csdms.colorado.edu), the EPA-funded CMAS (Community Modelling and Analysis System - http://www.cmascenter.org/), the DoD-funded CSTM (Community Sediment Transport Model - http://woodshole.er.usgs.gov/ project-pages/sediment-transport/) and others.

2. INTEROPERABILITY IN ENVIRONMENTAL OBSERVATORIES, AND COMMUNITY BUILDING

Our experience developing data infrastructure and modelling framework highlighted the following issues which are critical for successful model development and environmental forecasting, and – as we believe – reflect general deficiencies in data integration frameworks for environmental observatories. These include data availability, metadata and catalogues, differences in information models and data access protocols, differences in model requirements with respect to available data, differences in semantics among datasets, and difficulties in consensus building. These are each discussed briefly in turn.

2.1. Data availability, metadata and catalogues

Creating easy-to-use, uniform and scalable data and services publication and discovery mechanisms, and helping community members familiarize themselves with the available resources, are the basis for engaging community members into an efficient interoperable network. At the moment, the many environmental stakeholders maintain their own data archives and data access systems. For the Chesapeake Bay, for example, several community resource repositories and discovery interfaces are being developed. They include:

- The Chesapeake Information Management System (CIMS), which brings together multiple datasets assembled by federal, state and local agencies. CIMS datasets have consistent metadata descriptions based on FGDC Content Standard for Digital Geospatial Metadata and the metadata content is searchable via text search, with many datasets available for download (the search interface is at [29]).

- The GEON-based CBEO portal, developed within the on-going NSF-supported CEO:P project. Within this portal, users can publish their datasets and services and search for registered resources of different types, including shapefiles, raster images, Excel spreadsheets, relational databases, WMS/WFS services, web services, documents, etc. A subset of CIMS water quality database is already registered through this portal [30]. Catalogue search is available via the portal, and via the GEONSearch web service.

- Hydrologic Information System (HIS Server) being developed within the NSF-supported CUAHSI HIS project, which supports publication, discovery and retrieval of observations data via WaterOneFlow web services [31]. The services follow CUAHSI WaterML [6] specification which has been adopted as an Open Geospatial Consortium's (OGC) Discussion Paper. The data discovery services available within the HIS Server are tuned to observations data series [32] and support data search and retrieval across data repositories via desktop and online (Data Access System for Hydrology, DASH [33]) interfaces. Under the aegis of the NSF WATERS program, the software is now installed at ten WATERS Hydrologic Observatory Testbed sites across the country, three of which are affiliated with the Chesapeake Bay program [35]. In addition, a HIS Server node is established for the CBEO project, and available within the GEON-based portal mentioned above.

Despite recent integration efforts, such as making the CIMS data, and the HIS Server for the Chesapeake Bay, available via the GEON-based portal, the above information management frameworks remain largely disconnected. Additional work is required to reconcile metadata and information discovery protocols across the repositories, to enable users query and explore available data for the area regardless of the repository and the adopted data management framework. Such disconnect is indicative for the current state of environmental observing systems, and should be addressed via the creation of a common set of re-usable cyberinfrastructure modules supporting integration of disparate data and models.

2.2 Differences in information models for observations data, and data access protocols

Our earlier attempt to import CIMS observations data into CUAHSI ODM (Observations Data Model, [36]) uncovered several differences in metadata used by both systems. In turn, both information models are different from the Open Geospatial Consortium's (OGC) "Observations and Measurements" [37] model and their Best Practices specification for Earth Observation Products [38].

In addition, each of the nationally hosted environmental data sources, such as hydrologic data repositories at USGS and EPA, have different data access interfaces. The CUAHSI WaterOneFlow services provide a simplified, consistent way of accessing data from a combination of these sources. While similar in approach to the OGC web services specifications, the CUAHSI web services are not OGC compliant at the moment, though initial harmonization steps are outlined in the CUAHSI WaterML specification [39]. As a result users wishing to access these sources can only do so using CUAHSI-compatible client software. Other non-OGC, non-CUAHSI data servers require a still greater degree of data customization software. Other domains than hydrologic data have similar issues, and may not even have web-based access.

It should also be mentioned that current approaches do not yet address the problem of model-data coupling which is likely to be necessary in the next generation of complex physical, ecological, and chemical environmental models. Furthermore it is not yet clear what will be the data needs of the next generation of modelling tools. It is likely that at least some clarity on this issue will come from the Environmental Observatory initiatives currently underway.

2.3 Differences in model needs with respect to the available data

Models require increasingly large volumes of input data, which raises a performance problem for accessing the data via web services during model runs. It is necessary in many cases to download these large volumes of data from the national portals and transform the data in certain ways to prepare for use with the desired numerical or analytical models. Spatial data may be initially available as irregular grids, regular grids, or vector form such as points, lines, and polygons, while the model may only work with the data in a different form. The scale or resolution of the initial data may not be appropriate for a given model execution, and must be transformed to the correct scale. The temporal resolution and reference system of time-series observations can differ from that expected or needed by models. Each model may have its own specific data requirements, resulting in increasingly painstaking manual effort to locate, obtain, and transform the data in preparation for modelling applications. At a minimum a new data model is needed that reflects the relationship between modelling and current and future data streams. It is clear that the data and modelling process must be considered together if progress is to be made on either front.

Participatory Modelling [40-44] is becoming a recognized approach to modelling complex systems for decision-making. However, there are no agreed standards and platforms for data sharing and group model development available so far.

2.4 Differences in semantics between datasets available for different times and assembled in different research domains.

Another significant issue is the difference in nomenclatures and semantics among data sources for a given type of data, usually as a function of the compilation or hosting organization. The simplest differences may be in terms of language, keywords, metadata tags, etc., especially across subject domains and time periods. For example, the depth of a stream may be referred to as "gage height," "stage," or "waterlevel." A similar example is that the classification codes (identifiers) for variables such as nutrients (e.g., nitrogen) will vary, depending on the data producer's conventions. Semantic reconciliation is a well-recognized component of the data interoperability challenge [37, 45, 46, 47, 48, 49, 50, 51, 52, etc.]. Within both GEON and CUAHSI HIS, several technical approaches to ontology management and semantics-based integration have been explored. Beyond the technical issues, significant effort is required to make common nomenclatures and ontology translations accepted as the basis for community lingua franca.

2.5 Difficulties in consensus building.

The realization that interoperability is desirable and important is relatively recent. Most of the interoperability difficulties just described have arisen due to institutional and cultural conventions, rapid advance in open software, sensors and communication, all of which have evolved over long periods of time, from independent research communities, and without the priority among data producers for the need or benefits to coordinate efforts.

It is clear now that interoperability is not just a technical or technological problem, but also one of building trust and consensus within and between subject domain communities. This is not an activity that will arise on its own within any single agency, but requires a determined effort at multiple levels—from the leaders to the workers, across numerous organizations—to develop consensus-based approaches. Consortia such as the W3C (World Wide Web Consortium), OASIS (Organization for Advancement of Structured Information Systems), OGC, and many others have emerged over the last three decades to develop tools and techniques for interoperability at the level of basic information exchange mechanisms using the Internet. At the same time, consortia of scientific research centres have emerged, such as NCAR (National Centre for Atmospheric Research), EarthScope, CENS (Centre for Embedded Networked Sensing), NEON, GEON, CUAHSI, CBEO, and many others dedicated to enabling integrative, interdisciplinary research. Each of these major centres seeks to achieve some degree of harmonization and standardization of information models, semantics, ontologies, and accessing methods across multiple domains. Various governance patterns and structures are now emerging: interoperability standards that support identifier governance, vocabularies and structural definitions, and are supported by protocols, persistence implementation, and vocabulary publication and governance. However, the growing complexity of the standards themselves raises still more issues in terms of consistent interpretation, implementation, and usage.

In addition, significant institutional and cultural issues still remain that limit the efficacy of interoperability technologies and policies.

- (Institutional) Agreements may be needed between data providers and data users, concerning the scope of use, limiting liability, establishing means of cost recovery for source data compilation, addressing privacy and security concerns, and otherwise establishing trust between parties.

- (Institutional/Cultural) In many cases, established policies and procedures inhibit or even preclude interoperability. This can happen when a given community has invested substantial time and effort in developing data models, semantics, and ontologies that are in conflict with related models from other communities, leading to a sense of competition between the communities rather than cooperation and collaboration.

- (Cultural) Interest and willingness to share data is a pre-requisite to achieving interoperability, but field scientists can often be reluctant to share their data in useful ways, even when required by the funding agency. Developing trust and motivating data sharing "in the proper spirit" may require a substantial investment of resources for outreach. Resistance to sharing will be exacerbated if this issue is discounted or ignored.

The most straightforward aspects of interoperability are in the design and development of technologies and practices for cyberinfrastructures supporting federated data sources linked to users through service-oriented architectures and web services. This approach serves a wide range of government, business, and scientific application areas, and is the subject of nearly all the current standards work in progress at OASIS, OGC, and other IT standards organizations. OGC has evolved a process for testbed pilot projects that has achieved high levels of sponsored participation and collaboration by government and industry data providers and users. As a result of OGC testbed projects which use rapid prototyping as a way to accelerate the design, testing, and adoption of interoperability tools and standards, numerous national agencies in the US, UK, Europe, and Australia now routinely require OGC specifications to be applied in RFPs for enterprise and federated geospatial information portals and other systems. This is an important step toward alleviating the institutional barriers to interoperability. By engaging both the leadership and the programming staff of government agencies and commercial entities in the activities of defining and adopting geospatial standards and best practices, many of the trust issues at the institutional level are addressed, and organizational policies evolve accordingly. Now the issues of most concern within the IT standards consortia are not so much about whether to share data, but how to do so while managing data security, intellectual property rights and licensing, personal privacy, cost recovery, and adequate functionality and performance of transactions for data creation and update.

Much work is also being done in core technology for bridging and mediating between different semantics and ontologies. The Geography Markup Language (GML) [63] was initially developed by OGC members in 2000, and has steadily improved to support the representation of essentially any type of geographic feature, but it is too abstract and general for direct use within a subject domain (user communities are encouraged to adopt an application schema based on some reasonable subset of GML to precisely characterize a given information model). Another broadly applicable standard more recently adopted has to do with webs of sensor data: SensorML [64], TransducerML [65], Sensor Observation Service [66], and other related specifications, were jointly developed by OGC and IEEE (Institute of Electrical and Electronics Engineers), and are beginning to be used in environmental observatory context.

The IT standards organizations have not, in general, been involved in the harmonization of semantics and ontologies within specific subject domains. Each distinct information community needs to develop its own information models and metadata catalogues, simply because these must reflect the deep science within each subject area. These developments, including GeoSciML (GeoSciences Markup Language, based on GML), EML (Ecological Metadata Language), CML (Chemistry Markup Language), WaterML (for hydrologic applications), and other domain-specific ML's, represent important steps in the maturing of community attitudes and understanding toward sharing information models and data in effective ways. At the same time, the choice, or development, of a relevant data model and markup language adequately describing the variety of data presentation and transformation needs of a research domain, remains a challenge. Furthermore, because these developments involve agencies, universities, and other research centres around the world, there is growing pressure to overcome many of the institutional and international barriers to interoperability, which have existed.

3. THE VISION AND THE APPROACH

Addressing the above issues requires a comprehensive approach to data and model interoperability in the course of community modelling effort. It has the following components:

- Understand the interoperability needs of the community in a participatory and collaborative effort;
- Develop a "common component" approach to data/model tools and architectures;
- Develop research scenarios that would benefit from interoperability, and build research groups around them (e.g., hypoxia). Build consensus about interoperability architecture and standards supporting these scenarios;
- Expand on environmental system observatory ontologies, in particular for mapping variables to concepts;
- Bring together modellers and data providers to agree on common access protocols, enabling models to automatically search for data needed and link to data servers to download the information required for modelling. Design data interoperability for model input/output to help link models;
- Provide an environment to support participatory modelling efforts;
- Design and implement a pilot version of software tools required for model data integration over the web;
- Leveraging existing data repositories, create an environment where researchers can discover, visualize and retrieve available information in a standard uniform manner; and
- Understand and apply more of the existing OGC standards and best practices.

4. TOWARDS AN INTEGRATED COMMUNITY MODELING INFRASTRUCTURE

There are two major complementary components in this kind of research: software research and development, and community building and integration. Both are equally important. Within the software-related category, there are a number of distinct types of components, interfaces, and web services, which can be considered.

4.1 Data availability, and harmonizing data discovery interfaces

We envision that easy query and browse access to a community information system with multiple community-contributed resources, coupled with the ease of publishing data of common interest, is a critical component of community infrastructure. The task of **unifying data discovery and exploring data availability** has the following sub-tasks:

1) Register community-generated observations datasets via existing online systems. Identify the most promising systems and focus on them.

2) Develop a uniform data discovery services system. Explore data discovery standards being adopted by the OGC. In particular, OGC has developed a set of Catalogue Services [53], common specifications for discovery, browse, and metadata query interfaces that support uniform search against heterogeneous data and service catalogues. OGC's most recent work has focused on the ebXML Registry Information Model (ebRIM) profile of the Catalogue services, as a way to accommodate various application schemas and coordinate reference systems.

3) Construct data availability research, and develop guides for data discovery and interpretation. Data discovery and integration are greatly aided by the abilities to explore and visualize large volumes of observations data and catalogues efficiently and in greater detail. Visualizing such dynamics, and interpreting it, is critical for better understanding of the entire observations history and geography for various regions.

4.2 Harmonizing information models and data access protocols

This will require development of translators, cross-walks, and other "infrastructure glue" for components supporting structural and semantic interoperability across data. This could include:

1) Exploring the differences and harmonizing proposed information models for observations data, including the OGC's "Observations and Measurements" [61] and the CUAHSI's ODM [36] and verify that the information models are complete for the observation datasets available. For example, the benefits of structural harmonization and crosswalks, can be illustrated by the Australian WRON project that assembled OGC-compliant WFS (Web Feature Service) services to expose observations data for New South Wales. Creating XSLT transformations that expose the Australian WFS services as CUAHSI WaterOneFlow services [54] allows one to quickly tune the CUAHSI HIS online data access system (DASH) to New South Wales hydrologic observations [55].

2) Developing converters and translation services for datasets and models. The CUAHSI WaterML development and its harmonization with the O&M specification is of great interest to OGC and the WRON project. The US Department of Energy, CCA program (Common Component Architecture) is a case in point.

3) Expanding the technologies developed for observations data within existing observatories, including biogeochemistry data, land use, atmospheric data, etc.. Arriving at data interchange standards and ontologies applicable across communities is a challenging task. It requires explicating and comparing semantic frameworks used in the neighbouring fields, outlining information models for commonly used data sets, standardizing data discovery and access mechanisms, and assessing common integration scenarios, e.g., in support of comprehensive modelling on watersheds and estuaries.

4.3 Matching model needs with the available data

By applying the same formalism that treats data sets as independent modules that can be accessed just like the modelling modules, we can integrate the data sets into the modelling system as well.

We need model pre-processors that automatically find model inputs from datasets registered to data catalogues, retrieve available data via web services, and assemble them in formats ready for modelling. Assembling model inputs is one of the most time consuming tasks in comprehensive environmental modelling, and requires knowledge of available datasets from several related domains (hydrologic, biological, atmospheric, social-economic, and demographic data, etc.). Making data discovery and assembly dynamic, even for a fraction of model inputs, has the potential of improving the efficiency and relevancy of environmental models, at the same time reducing proliferation of multiple versions of identical or almost identical datasets. For example consider the inputs and data flows for the CBP Phase 5 model [9,10] or similar data for PSU's PIHM model [11,12] (Figure 1).

The data structure should be designed to be flexible enough for modification and customization of the model, and rich enough to represent complex user defined spatial relations, and extensible to add more software tools as the need arises. For this we may need to explore new integration methodologies for linking a GIS framework with physical models that enables users to take advantage of object oriented programming (OOP) with direct access to the GIS data structure, which support efficient query and data transfer between the model and GIS.

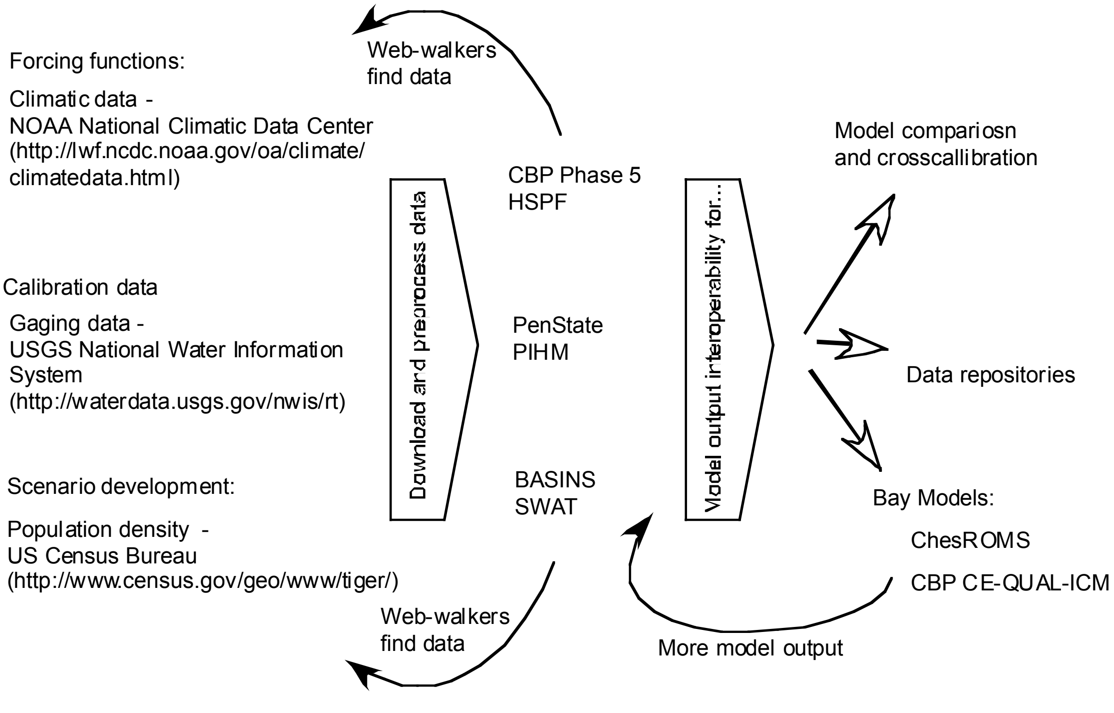

Figure 1. An example of data flow between some data sets and models. There are many more models and data sets, but these will be primarily targeted for this pilot study [11-25].

Suppose we have a watershed model such as the HSPF [7,8] model that is core of the Chesapeake Bay Program (CBP) [9,10]. Like other watershed models, it requires information about climatic conditions and flow data for streams. These data sets are available from the web; however, substantial effort is required to download all the information needed and convert it to the proper format for landscape modelling. Each time we move from one sub-watershed to another, this effort must be repeated.

It is essential to have the right software tools and consensus between data providers (in this case, NOAA and USGS) and modellers (in this case, CBP) to make sure that the watershed model can access the data needed as a standard pre-processing, setup routine, when the data will be found and downloaded for further model runs.

In addition to climatic and water data, the watershed models are linked to socioeconomic information required for landuse coverages and calculation of loading factors. This leads to further exploration of linkages to census data available from the US Census Bureau [56]. The census data is organized according to census blocks and tracts, which have nothing in common with the watershed and subwatershed spatial structures assumed in watershed modelling. We need additional preprocessing to reorganize and resample this data to make it available for the model.

Additionally, these standardized protocols for data access should be available to any models in the CCMP directory and beyond. For example, the PSU PIHM [11,12] watershed model has similar data requirements, but runs over a triangular spatial grid. Resampling procedures should provide similar data access for a different geometric structure used in this model. The output from these models can be then piped into Bay models such as the CBP-QUAL-ICM or ChesROMS, which is part of the CCMP open source distribution.

4.4 Ontology and semantics

Exploring differences in vocabularies for variable names, units, collection methods, sampling mediums, etc., used by local data collection agencies and projects, and developing semantic cross-walks, comprise a critical interoperability component for building community networks. Ontologies can be harmonized effectively via ontology-focused "variable tagging" workshops, in which measured variables are mapped to broader concepts, which in turn belong to even broader concepts eventually ending at the "top concept." The key to this effort is that at lower levels a certain degree of heterogeneity must be accepted because experience shows that it is extremely difficult to reach consensus on what to call a certain variable. It is more prudent (and in a way provides a much higher degree of harmony and as such acceptance) to permit variety at this level to obtain buy-in because a high degree of flexibility is maintained, but then merge different variables at higher concepts and use the higher (or more general concepts) as an entry for search queries. Furthermore, the effort must reflect the continuing advances in cyberinfrastructure tools and resources which will require a flexible or evolutionary approach if these advances are to reach the domain community. This work will leverage the initial set of tools for registering ontologies, ontology tagging, and ontology-based search developed within GEON and CUAHSI HIS. For hydrologic data, these tools include HydroSeek [57], an ontology-aided search engine developed at Drexel university that can search over USGS, EPA, and several local observation networks including the CIMS databases, and the ontology tagger tool [58] which is a graphical user interface that helps users associate variable names in an ODM database with concepts in a hydrologic ontology, to make the database available for ontology-based search. The underlying ontologies are currently focused on physical and water quality parameters that are of primary interest to hydrologists or environmental engineers, hence they need to be build up to encompass a larger spectrum that is also of importance to for example oceanographers and ecologists.

4.5 Serving data for participatory modelling efforts

In recent years, there has been a shift from top-down prescriptive management of water resources towards policy making and planning processes that require on-going active engagement and collaboration between stakeholders, scientists, and decision-makers. Participatory modelling is the process of incorporating stakeholders, often including the public, and decision-makers into an otherwise purely analytic modelling process to support decisions involving complex ecological questions. It is recognized as an important means by which non-scientists are engaged in the scientific process and is becoming an important part of watershed planning, restoration, and management. The development of unique, practical, and affordable solutions to ecological problems is often best accomplished by engaging stakeholders and decision makers in the research process. These group modelling efforts require specific types of models and data to be successful. These modelling tools are usually simpler than what we find in full-fledged research models, and for example might use Excel, Stella, MATLAB, or Mathematica as a means of joint learning and system representation. The scoping models that are produced are designed to gain shared experience about the system and build consensus among stakeholders. However, they also require data to make them run. Moreover, they need a lasting web presence that would support group interactions, and link directly to diverse data and modelling tools.

4.6 Community consensus-building and testbed processes

Various communities need to be engaged in as many ways as possible. This can be achieved by the following means:

1) On-going "cyber-seminars" for which selected participants would submit white papers in advance for consideration by the community. On specific days, discussions of the content could take place using simple web conferencing tools such as Skype, Webex, etc. Follow-up discussion would take place on a twiki over a prescribed period of time (one to two weeks). The results of discussion could be incorporated into the initial white paper(s) and re-posted to the community, which could then decide on subsequent action: e.g., allow the paper to form the basis for subsequent research and development work; redirect research currently underway; submit for presentation to a conference; etc. The twiki could be used for further feedback and subsequent results as needed.

2) Provide means for a Web 2.0 approach to obtain community rating and annotation of submitted resources. Each key resource, such as source data sets, derived data sets, metadata catalogs, software tools developed to search the metadata and actual data sets, etc., will have a commonly identified means in the portal for users to rate the resource and enter comments explaining their ratings. The comments are essential to enable the resource developers to understand and respond to any problems or other issues. These comments and follow-up will be posted to resource-specific twikis for tracking and archive.

3) Support Inter- and intra-community ontology development with a Web 2.0 approach, along the lines of urbandictionary.com. Terms of reference, classification systems, valid values for coded domains, etc., will be posted to twikis allowing all users to see and vote on competing definitions. As specific issues arise that require more extensive discussion for resolution, the community will be called together for a cyber-seminar.

Solving the cultural issues mentioned previously is more difficult than the institutional issues, because these very often come down to working with attitudes, opinions, and beliefs of strong-willed individuals who may resist external directives. But here too, lessons are being learned and progress is being made. An individual field scientist can hardly be faulted for reluctance to share her data when she doesn't know the motives and applications a given user might have in mind. She may not see or accept the value to her of allowing her data to be accessed and used without direct involvement. But if, say, someone from the cyberinfrastructure community were to take a personal interest in the work, and earn trust based on a relationship of shared understanding, then the perspective is more likely to open. One needs to be able to see and trust ways of benefiting from the synergistic implications of contributing one's data to a greater body of knowledge. (It is also likely that some form of external rewards could help to further encourage this outcome.) Certainly, even with personal attention and effort to create such bridging relationships with the IT community, some scientists will still reject the importance of sharing their data, and of working with others to use community-based information models and data exchange mechanisms. This is just part of human nature. Such scientists may yet respond to firm directives from funding agencies to adhere to community information models and best practices, but the difference between motivated and unmotivated adherence to recommended standards and practices can make a big difference in the fitness of data for other uses. It should also be mentioned that funding resources and funding agencies should better reflect the changing environment described above, and institutions must find new metrics for evaluating excellence of scientists who commit to community data and model development.

In summary, while the focus to date on interoperability standards and tools has been largely technical, we need to acknowledge and work with the social aspects of it now. In order to move forward, we must now focus on forming consensus around information and modelling requirements and architectures, making the differences between current information models and workflows explicit, helping to build solid relationships between IT and subject domain scientists, and otherwise engaging the various stakeholder communities in these activities as much as possible.

REFERENCES

[1] Collaborative Large-Scale Engineering Analysis Network for Environmental Research (CLEANER): An Engineering Cyberinfrastructure "Test Bed" (CLEANER), project office (http://cleaner.ncsa.uiuc.edu/home/)

[2] Consortium of Universities for the Advancement of Hydrologic Science, Inc. (http://www.cuashi.org)

[3] National Ecological Observatory Network (http://neoninc.org/)

[4] GEON; The Geosciences Network (http://www.geongrid.org)

[5] The Long Term Ecological Research Network (http://www.lternet.edu/)

[6] The WATer and Environmental Research Systems Network (WATERS Network) (http://watersnet.org/)

[7] Donigian, A.S., J.C. Imhoff, B.R. Bicknell, and J.L. Kittle. 1984. Application Guide for Hydrological Simulation Program - FORTRAN (HSPF). Athens, GA, U.S. EPA, Environmental Research Laboratory.

[8] Donigian, A.S., Jr., B.R. Bicknell, and J.C. Imhoff. 1995. Hydrologic Simulation Program - FORTRAN (HSPF). Chapter 12 in Computer Models of Watershed Hydrology, V.P. Singh, Ed., Water Resources Publications, Littleton, CO.

[9] Donigian, Jr., A.S., B.R. Bicknell, A.S. Patwardhan, L.C. Linker, C.H. Chang, and R. Reynolds. 1994. Chesapeake Bay Program Watershed Model Application to calculate bay nutrient loadings. U.S. EPA Chesapeake Bay Program Office, Annapolis, MD.

[10] http://www.chesapeakebay.net/temporary/mdsc/community_model/about.htm

[11] Duffy, C.J. 2004. Semi-discrete dynamical model for mountain-front recharge and water balance estimation, Rio Grande of southern Colorado and New Mexico, Pages 255-271 in: J.F. Hogan, F.M. Phillips, and B.R. Scanlon (eds.), Groundwater Recharge in a Desert Environment: The Southwestern United States. Water Science and Applications Series, vol. 9, American Geophysical Union, Washington, D.C.

[12] Qu, Y., and C. J. Duffy (2007), A Semi-Discrete Finite-Volume Formulation for Multi-Process Watershed Simulation, Water Resour. Res., doi:10.1029/2006WR005752

[13] http://www.epa.gov/waterscience/basins/index.html

[14] http://ccmp.chesapeake.org/CCMP/models/ChesROMS/index.php

[15] Cerco, C., and Cole, T., 1994: Three-dimensional eutrophication model of Chesapeake Bay, Technical Report EL-94-4, US Army Engineer Waterways Experiment Station, Vicksburg, MS.

[16] http://el.erdc.usace.army.mil/elmodels/icminfo.html

[17] http://www.ccalmr.ogi.edu/CORIE/modeling/elcirc/

[18] Zhang, Y.-L., Baptista, A.M. and Myers, E.P. (2004) "A cross-scale model for 3D baroclinic circulation in estuary-plume-shelf systems: I. Formulation and skill assessment". Cont. Shelf Res. 24: 2187-2214.

[19] Cerco, C.F., 1995: Simulation of long-term trends in Chesapeake Bay eutrophication, Journal of Environmental Engineering, vol. 121(4), p. 298-310.

[20] Sheng, Y.P., 1986: A Three-dimensional Mathematical Model of Coastal, Estuarine and Lake Currents Using Boundary Fitted Grid, Report No. 585, A.R.A.P. Group of Titan Systems, New Jersey, Princeton, NJ, 22 p.

[21] http://www-nml.dartmouth.edu/circmods/gom.html

[22] http://chartmaker.ncd.noaa.gov/csdl/op/c3po.html

[23] Xu, J. and R.R. Hood. (2007). Modeling biogeochemical cycles in Chesapeake Bay with a coupled physical-biological model. Coastal Shelf Est. Sci.

[24] http://www.hydroqual.com/wr_rca.html

[25] Di Toro, D.M. 2001. Sediment Flux Modeling. J. Wiley and Sons. New York.

[26] Acock, B. and J. F. Reynolds (1990). Model Structure and Data Base Development. Process Modeling of Forest Growth Responses to Environmental Stress.. R. K. Dixon, R. S. Meldahl, G. A. Ruark and W. G. Warren. Portland, OR, Timber Press.

[27] http://wise2.jrc.it/wfdview/php/index.php

[28] Baru, C. (2005) GEON: The GEON Grid Software Architecture. ESRI International User Conference, San Diego, August 2004.

[29] http://www.chesapeakebay.net/data/index.cfm

[30] http://geon16.sdsc.edu:8080/gridsphere/

[31] http://www.cuahsi.org/his

[32] http://river.sdsc.edu/wateroneflow/

[33] http://river.sdsc.edu/DASH

[34] http://watersnet.org/wtbs/wtbs08/index.html

[35] http://watersnet.org/wtbs/wtbs10/index.html

[36] CUAHSI Observations Data Model (ODM), (http://www.cuahsi.org/his/odm.html)

[37] Ouskel, A. and A. Sheth. 1999. Semantic Interoperability in Global Information Systems. ACM SIGMOD Record Vol. 28 No. 1 pp. 5-12.

[38] GML 3.1.1 Application schema for Earth Observation products, 2007. Editor Jerome Gasperi, OGC Best Practices Document, OGC 06-080r2 (http://portal.opengeospatial.org/files/?artifact_id=22161).

[39] [WaterML07] CUAHSI WaterML, Editors Ilya Zaslavsky, David Valentine, Tim Whiteaker, 2007, OGC Discussion Paper, OGC 07-041r1 (http://portal.opengeospatial.org/files/?artifact_id=21743)

[40] Brown Gaddis, E. J., Vladich, H., and Voinov, A. 2007. Participatory modeling and the dilemma of diffuse nitrogen management in a residential watershed. Environ. Model. Softw. 22, 5 (May. 2007), 619-629.

[41] A. Castelletti and R. Soncini-Sessa, 2006. A procedural approach to strengthening integration and participation in water resource planning Environmental Modelling & Software, Volume 21, Issue 10, Pages 1455-1470

[42] Joanne Tippett, 2005. The value of combining a systems view of sustainability with a participatory protocol for ecologically informed design in river basins Environmental Modelling & Software, Volume 20, Issue 2, Pages 119-139

[43] Roberts, N., 2004. Public deliberation in an age of direct citizen participation. American Review of Public Administration 34(4), 315-353.

[44] Korfmacher, K. S., 2001. The politics of participation in watershed modeling. Environmental Management 27(2), 161-176.

[45] Bergamaschi, S. Castano, S. and Vincini, M. 1999. Semantic integration of Semi-structured and Structured data Sources. Special Issue ACM SIGMOD Record Vol 28. No; 1 pp. 54-59.

[46] Bishr Y. 1998. Overcoming Semantic and other barriers to GIS interoperability. International Journal of Geographic Information Science. Vol. 12 No.4. pp. 299-314.

[47] Castillo, J. A., Silvescu, D. Caragea, J. Pathak, V. Honavar. 2003. Information Extraction and Integration from Heterogeneous, Distributed, Autonomous Information Sources: A federated Ontology-Driven Query-Centric Approach. IEEE International Conference on Information Reuse and Integration (IRI 2003) Las Vegas, NV.

[48] Garcia-Solaco, M., Saltor, F. and Castellanos, M. (1996). Semantic Heterogeneity in Multidatabase Systems, in Bukhres, O. and Elmagarmid, A (Eds) Object-Oriented Multidatabase Systems. Prentice Hall: Englewood Cliffs, NJ. pp 129-202.

[49] Kashyap, V. and Sheth, A. (1998). Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In Papazoglou, M. P. and Schlageter, G., editors, Cooperative Information Systems, pages 139--178. Academic Press, San Diego.

[50] McLeod, D. Fang, D. and Hammer, J. 1991. The Identification and Resolution of Semantic Heterogeneity. In Proceedings of First International Workshop on Interoperability in Multidatabase Systems. Kyoto, Japan, April 1991.

[51] Shahabi C. and Sheth, A. 1991. Semantic Issues in Multidatabase systems. SIGMOD special issue. Vol 20. No. 4.

[52] Welty, C. and N. Guarino 2001. Supporting Ontological Analysis of Taxonomic Relationships. Data and Knowledge Engineering, 39 (1), 51-74

[53] http://www.opengeospatial.org/standards/cat

[54] http://river.sdsc.edu/wiki/CSIRO%20Collaboration.ashx

[55] http://river.sdsc.edu/awdip/

[56] http://www.census.gov/geo/www/tiger/

[57] http://www.hydroseek.org

[58] http://water.sdsc.edu/tagger

[59] http://www.ogcii.org/bod

[60] Chesapeake Bay Environmental Observatory
(http://ccmp.chesapeake.org/CBEO/resources.php)

[61] Observations and Measurements, 2006. Editor: Simon Cox, OGC Best Practices Document, OGC 05-087r4 (http://portal.opengeospatial.org/files/?artifact_id=17038).

[62] http://www.opengeospatial.org/standards/gml

[62] http://www.opengeospatial.org/standards/sensorml

[63] http://www.opengeospatial.org/standards/tml

[64] http://portal.opengeospatial.org/files/?artifact_id=12846