Proceedings of the Institution of Civil Engineers
Water Management 163
January 2010 Issue WM1
Pages 13–25
doi: 10.1680/wama.2010.163.1.13

Paper 800055
Received 05/05/2008
Accepted 14/01/2009

Keywords: hydrology & water resource/information technology/ knowledge management

Marian Muste
Research Engineer, University of Iowa, Iowa City, IA, USA

Dongsu Kim
Postdoctoral Research Associate, University of Iowa, Iowa City, IA, USA

Nicholas Arnold
Graduate Research Assistant, University of Iowa, Iowa City, IA, USA

Tim Whiteaker
Research Associate, University of Texas at Austin, Austin, TX, USA

Craig Just
Research Engineer, University of Iowa, Iowa City, IA, USA

Anton Kruger
Associate Professor, University of Iowa, Iowa City, IA, USA

Digital catchment inception using community project components

M. Muste PhD, D. Kim PhD, N. Arnold BS, T. Whiteaker PhD, C. Just PhD and A. Kruger PhD

It is becoming recognised that synergistic integration of knowledge and methods from a variety of allied disciplines is needed to study, engineer and manage water resources. Until recently this comprehensive investigative approach was limited, in part due to a lack of appropriate technologies. Recent advances in computer, communication and information technologies have led to increasingly pervasive and sophisticated 'cyberinfrastructure', which enables integration in a common digital environment using remote and in situ measurements with geotemporal databases and high-performance computational models. This paper describes the initial efforts carried out by a University of Iowa research group in implementing and customising software components created by a hydrologic community project to support a local digital catchment. The initial goal was to test what types of data queries the digital catchment can handle and to see how it performs in cases where data streams are coupled with models for continuous forecasting of river water quality. The paper also discusses the general context for digital catchment development and summarises the lessons learned during this initial developmental stage. Given the uniform and scalable nature of the community project components, the workflows described are transferable to other users and other catchments. This synergistic initial effort led to the conclusion that community project components can successfully underpin the national effort of creating a network of ecohydrologic observatories.

1. INTRODUCTION

There is substantial concern that peoples' water usage may significantly alter the variability and evolutionary trajectory of water-cycle systems at local and regional scales. Anthropogenic activities interfere with natural procedures to alter processes in these catchment systems dramatically – and often irreversibly. Although many studies have examined human–water dynamics (e.g. Liu *et al.*, 2007), the complexity of such coupled systems is not completely understood because

(*a*) there are gaps in understanding of water-centric bio-geochemical and socio-economic processes
(*b*) analyses are still strongly disciplinary and inward looking
(*c*) appropriate tools and data for multi-disciplinary studies do not exist.

The complexity of these systems stems from their non-linear dynamics, scale-dependent behaviour, reciprocal feedback loops, time lags, resilience and heterogeneity of the interacting processes (Liu *et al.*, 2007). Addressing these issues requires a substantially better understanding of the linkages and feedback between the various systems than is presently available (Hall, 2003; Hall and Anderson, 2002; Price, 2000). Hydroscience and environmental engineering communities are responding to these challenges with a new investigative approach that takes into account all water-cycle components and their interactions with ecological, bio-geochemical and human systems at the catchment-scale level. The new approach is enabled and sustained by the fast evolution of digital-based remote and in situ observation and communication technologies coupled with the assimilation of geographic information system (GIS)-based relational databases and high-performance computing. Collectively, these advancements have led to an 'information-centric' approach for catchment investigation and management that capitalises on observations and their interpretation (Maidment, 2008).

In the USA, current efforts at improving the infrastructure and methodologies for integrated water-centric studies are being led by two relatively new National Science Foundation (NSF) communities

(*a*) the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (Cuahsi)
(*b*) the Water and Environmental Research Systems Network (Waters Network).

Cuahsi and Waters Network have promoted alliances among US universities to develop and implement cyberinfrastructure-based environmental observatories for examining catchments. 'Cyberinfrastructure' (CI) is a new term coined by the NSF to define the coupled use of real-time sensing (temporally and spatially detailed), databases, high-performance computational platforms and computer models to enable understanding of processes, knowledge management, visualisation, interaction and collaboration in all science and engineering disciplines (NSF, 2008). CI for water sciences is the American homologue for hydroinformatics (Abbott, 1991). The CI-based observatories promoted by Cuahsi and Waters Network seek innovative and effective approaches and methodologies to illuminate the processes governing catchment environmental

Institution of Civil Engineers
IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

systems, thereby greatly aiding catchment management, restoration and resource optimisation. A national network of observational and experimental facilities for systematic water measurements, data storage and curation, modelling and visualisation is envisioned that will enable unprecedented scientific and engineering research (Waters Network, 2008a).

The scientific communities have made great strides in clarifying frameworks and terminology and developing conceptual models for CI-based environmental observatories. Observatories deployed in natural environments are not unique to the US or world scientific communities. A number of similar initiatives have been launched in the USA in the past decade, for example the National Ecologic Observatory Network and the Geosciences Network. In the international arena, the World Meteorological Organization plays a leading role in developing information systems at the global scale and oversees a set of global observing systems (Global Ocean Observing System, Global Atmosphere Watch, World Hydrological Cycle Observing System). The observatories' primary role is to enable full-scale hypothesis testing and verify pilot-study findings through observations at larges scales in comparable settings (Waters Network, 2008a). Observatories enable cross-disciplinary process discovery and understanding through observation, simulation, analysis and knowledge synthesis using components and connections as illustrated in Figure 1 (Muste, 2007). They contain experimental facilities strategically implemented within the observatory. Simulation models are used to understand processes, to support the observatory design and to predict process aggregation at larger scales using various scaling theories. Critical CI is needed to enable communication, storage and knowledge analysis acquired from the deployed sensing systems and other available data sources, as well as modelling, collaboration and knowledge networking support.

Central to the observatory engineered system is the digital catchment (DW) concept; this is a comprehensive characterisation of ecohydrologic systems using an object-oriented representation of the catchment characteristics, behaviour and relationships as obtained from observations and numerical simulations (i.e. streams, lakes, hillslopes, etc.) to facilitate the study of multi-scale, multi-process dynamics of catchments (Maidment, 2008). For systems as complex as catchments, DWs hold the greatest promise for advancing insight and management. The DW description covers both the natural environment and manufactured infrastructure (e.g. dams, water abstraction and discharge systems). A comprehensive DW must embrace the best available information produced from all sources, which may include data and simulation models produced by various stakeholders (e.g. federal, state and local agencies, water authorities and districts, cities, counties, consultants). It contains the means to track the movement of water, sediments, contaminants and nutrients through the environmental system. The DW should support

(*a*) querying catchment-related information and behaviour at various levels of complexity

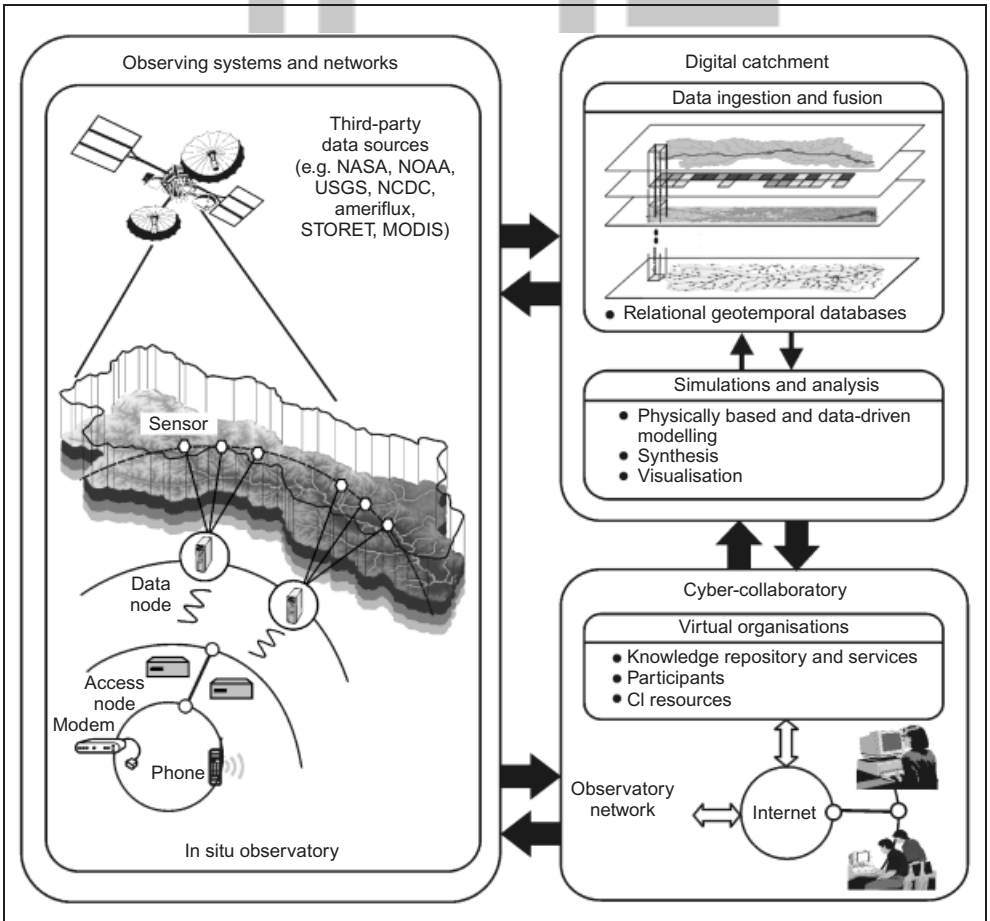

Figure 1. Conceptual configuration of the CI-based environmental observatory

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

(*b*) assessing DW consistency and completeness
(*c*) publication of DW data as web services
(*d*) instantiating a DW (populating the data model) to obtain a data warehouse for further off-line analysis.

A DW must be interoperable with a range of other observatories and be scalable (change in scale, component location, data types and community processes, scientific understanding, incoming technologies).

There are a number of critical obstacles to overcome in developing local DWs. Among the challenges are

(*a*) agreement on common ontologies, vocabularies and protocols
(*b*) engagement in a vast number of distributed data resources regardless of their provenance
(*c*) simple data discovery systems with distributed and integrated catalogues
(*d*) development of multi-task web services for automated real-time data handling.

These challenges are currently being addressed for hydrology by the multi-disciplinary research teams grouped under Cuahsi's hydrologic information system (Cuahsi-HIS) community project (Cuashi-HIS, 2008). The project has already produced tools and procedures that can be implemented by closely following Cuashi-HIS user manuals. However, development of customised workflows using Cuahsi-HIS components can be difficult for domain scientists as some knowledge of computer science and information systems is required to configure them correctly.

The purpose of this paper is to share the experience gained during the initial steps of building customised applications for a local observatory that couple Cuahsi-HIS tools and functions with the data and data model of a local catchment. While each of the workflows presented here was developed for specific purposes, the customisation process is quite generic and readily adaptable to other observatories. Section 2 summarises the main Cuahsi-HIS project components and their role. Initial workflows customised for a regional digital observatory located in Clear Creek, Iowa, are in Section 3 to illustrate the scalability and portability of Cuahsi-HIS components. In this initial developmental stage, local observatory efforts tested what types of data queries the system can handle and how it performs in cases where data streams are coupled with models for continuous operation. The paper then discusses lessons learned during the initial implementation stage and how the assembled components can benefit scientific hypothesis testing.

2. CUAHSI-HIS SHARED COMPONENTS

Originating in 2004, the Cuahsi-HIS is a geographically distributed network of hydrologic data sources and functions that use web services in order to function as an integrated system (Cuashi-HIS, 2008). The goals of the Cuahsi-HIS project are to

(*a*) unite national water information
(*b*) make this information universally accessible and useful
(*c*) provide access to the data sources, tools and models that enable synthesis, visualisation and evaluation of the behaviour of hydrologic systems.

Recently, specialised subgroups of Waters Network reviewed the Cuahsi-HIS approach, anticipating that the system might be compatible with the aspirations of both institutions, and consequently evaluated the possibility of extending the HIS cyberenvironment for a variety of environmental data (Waters Network, 2008a). The Cuahsi–Waters Network joint efforts target creating a national-scale network of observatories. For this purpose, 11 local observatories were deployed (see Figure 2) to demonstrate that water observation data collected by academic investigators could be

(*a*) stored in a standard way in a relational database
(*b*) published on the internet
(*c*) federated with water observation data published by water agencies
(*d*) searched using a concept framework that connects with variables in each individual data source (Goodall *et al.*, 2008).

A wide variety of comprehensive documents and other publications describing the tools and functions entailed by the Cuahsi-HIS project is available (Cuashi-HIS, 2008). This section briefly describes the salient components and system features to provide the context for the local observatory implementation and to delineate the contributions of the present research team from those of the community project. The Cuahsi-HIS software stack architecture is schematically shown in Figure 2. The Cuahsi-HIS server leverages commercial-off-the-shelf (COTS) software (e.g. SQL Server 2005, ArcGIS Server) to store, analyse and publish hydrologic observation data; it also includes free data storage and data publication components developed by the Cuahsi-HIS team.

The central system component is the observations data model (ODM). This provides a common database structure in which syntactically and semantically mediated observations – regardless of source, collection method or original file type and format – are stored along with their metadata (Horsburgh *et al.*, 2008). The ODM stores these observations into an SQL Server database. Each sensor or instrument network can have its own ODM database and GIS data layer, a network being defined as a set of sites with observations of variables that logically belong in the same database (Whitenack, 2007) (e.g. all of the US Geological Survey (USGS) gauges measuring real-time stream flow are grouped into a single observations network). The ODM includes time series values and timestamps, along with value-level metadata such as offset, method, quality control level and data qualifier. The ODM also includes several metadata tables that provide information about sampling sites, variables, source organisation and data provenance. At this time, the ODM only handles scalar point data sources.

While the ODM provides an effective database schema for storing data, scientists must still transform raw data into the ODM format. To assist with this task, the HIS server includes two data loaders: the ODM Data Loader (ODMDL) and the streaming data loader (SDL). The ODMDL is designed for manually importing data from a comma-delimited file or spreadsheet. The SDL is designed to work in real time with

Institution of Civil Engineers
IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

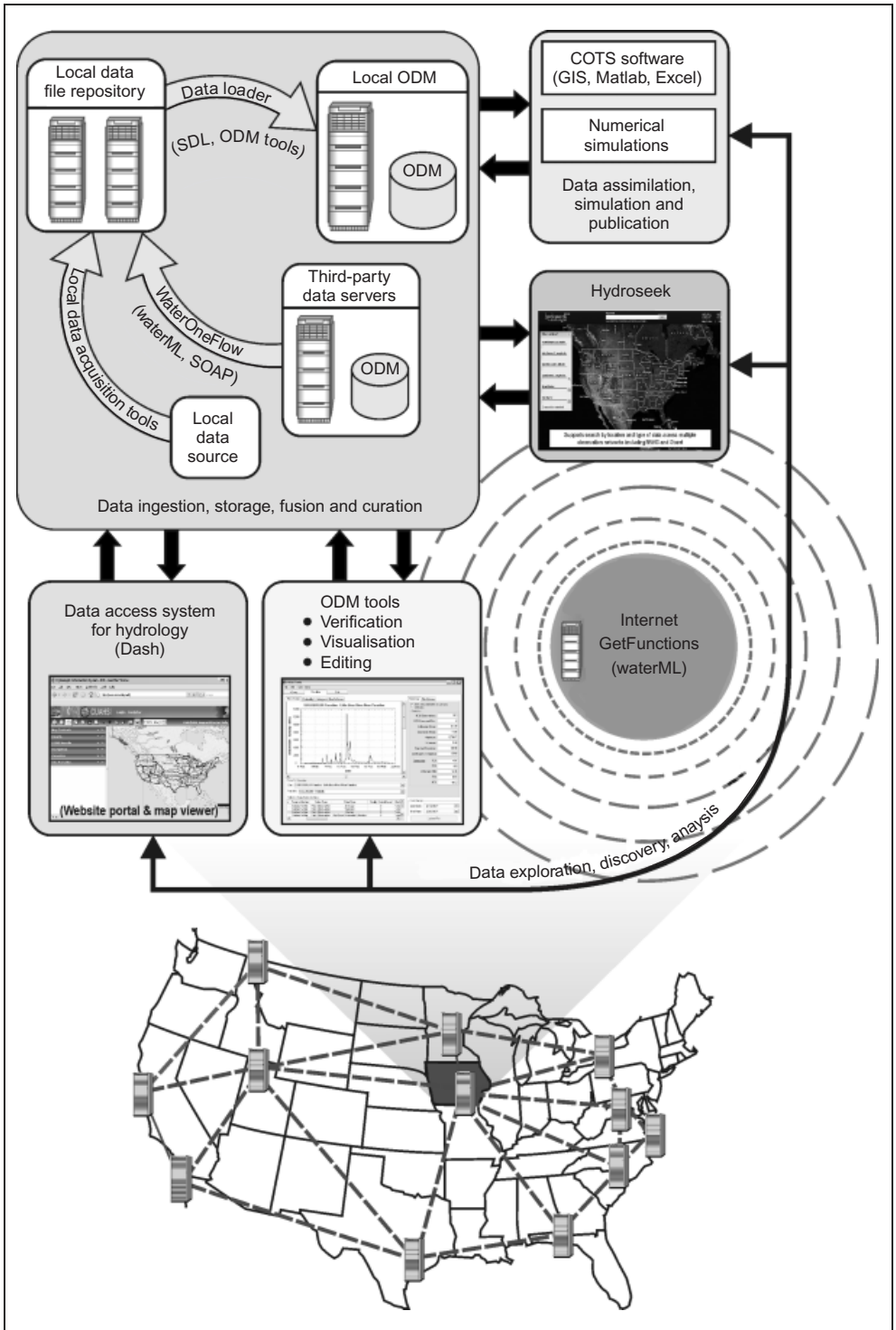

Figure 2. Main components of the Cuahsi-HIS and the network of servers deployed at Waters Network test bed sites (Maidment, 2008). Each server has developed and deployed an ArcGIS server uniformly equipped with Cuahsi-HIS components

incoming files from dataloggers such as those generated by Campbell Scientific's LoggerNet software. Since data are often received as streaming data files from sensors, the SDL is designed to run periodically so that the ODM database will always have the most current sensor data. Once the data are loaded into an ODM database, data analyses can be performed with the ODM Tools software (Horsburgh, 2007). ODM Tools enables a user to query an ODM database for a data series that matches a number of search parameters, plot graphs from a searched data series, edit data points and derive new data series. The ODM Tools software can also export a queried data series to common formats such as comma-delimited files (CSV).

While an academic investigator's data will no doubt be useful for their own analyses, they are often useful for projects conducted by other scientists. Therefore, the Cuahsi-HIS server includes technology for publishing data over the internet or local computer network using Cuahsi's WaterOneFlow web services. This provides a uniform method for querying data from a given data provider (Whiteaker, 2007). Web services are

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

applications that provide the ability to pass information between computers over the internet in a standardised format. The WaterOneFlow web services transmit data extracted from an ODM database encoded as extensible markup language (XML) and formatted using an XML schema called WaterML (Valentine *et al.*, 2008). In other words, what goes into a given data provider as a query is always structured the same way. This means that a user would ask for data from a given academic investigator's system in the same manner as asking for data from numerous other investigators or even national agencies such as the United States Geological Survey (USGS) who have also implemented WaterOneFlow web services via San Diego Supercomputer Center (SDSC).

However, while the generic WaterOneFlow web service makes a scientist's observation data accessible via the internet, the data can still be hard to access for a user who is not familiar with using web services. Cuahsi-HIS therefore developed several tools that build off WaterOneFlow web services to add value to the data that these web services provide. One of the most obvious needs is a web-based user interface for working with those web services. The Cuahsi-HIS data access system for hydrology (Dash) fills that role by letting a user search for and download hydrologic observations data using a map of observation sites in a web browser (Maidment *et al.* 2007). Dash utilises WaterOneFlow web services to provide users with the requested information. As long as a given web service provides data in WaterML format, it can easily be hooked up to Dash. Once this task is accomplished, the scientist now possesses a powerful server whose components all operate in synchronicity to support data storage, analysis and publication.

3. BUILDING THE INITIAL CLEAR CREEK DW WORKFLOWS

3.1. Overview of local context

For the past four years, an on-campus interdisciplinary group has been researching how the NSF CI initiative could be applied to environmental research at the University of Iowa (UI). The informal group, called CyberEnviroNet (UI, 2008a), comprises more than 25 scientists and engineers from several UI departments (engineering, geography, geoscience, computer science). Since its inception, an important target of the research group's efforts has been the development of a DW for Clear Creek (CCDW) with the aim of supporting understanding of catchment processes and potentially providing a scientific basis for decision-making in water resources. The first prototype of the CCDW, established in 2005, was developed using the Arc Hydro data model (GIS WRC, 2008) along with customised software components developed in-house for database extension, data retrieval and manipulation (Muste *et al.*, 2006). In 2006, the Clear Creek catchment became one of the 11 Waters Network test beds serving as beta test locations for the deployment of the Cuahsi-HIS project products (UI, 2008b). Consequently, the UI research team transitioned developmental efforts toward Cuahsi-HIS products to take advantage of the new server and ancillary software components provided by the national project for the local research interests.

The CI-based observatories promoted by Cuahsi and Waters Network seek innovative and effective approaches to illuminate processes governing catchment environmental systems, thereby greatly aiding catchment management, restoration and resource optimisation. It is anticipated that a national-scale network of observational facilities will be established in order to enable unprecedented science and engineering research. Investigators in local observatories will participate in the development and deployment of a common hydrologic information system that enables cross-domain analysis within individual test beds as well as cross-test-bed analysis and sharing of data. The concept behind the network is that science will benefit from greater access to data. For example, the network can

(*a*) explore one environmental variable while keeping other variables constant
(*b*) compare the same data measured in the same way at multiple sites
(*c*) facilitate understanding of broad water-driven patterns of behaviour, therefore generating interdisciplinary understanding and collaboration.

The initial developmental task for CCDW was the integration of simple but diverse sensors and communication means with data and numerical models into an end-to-end system that can operate via the internet automatically and in real time, as summarised in Figure 3. Currently, the CCDW handles the data streams listed in Table 1; details on third-party sources of data and the in situ instruments and associated data communication are discussed by Just *et al.* (2007). The data can be queried, downloaded or visualised using the Dash web application's built-in functionality. The next task was to ingest data streams in a simple water-quality simulation model for forecasting information related to the safety of the stream at the catchment outlet (see warning flags in Figure 3). The above applications are customised workflows that make extensive use of components developed by the Cuahsi-HIS project.

The first step for a new user of Cuahsi-HIS products is to become familiar with the training materials made available for the community (Cuashi-HIS, 2008). Using the guidelines, a user can configure the Cuahsi-HIS software set (i.e. Dash, WaterOneFlow, ArcGIS Server, SDL) to produce some initial results such as making data available through Dash and the web services. Extension of the system functionality through customised workflows requires the user to be comfortable working with a number of technologies (including ArcMap, ArcGIS Server, XML, .Net, ASP 2·0, Ajax) and web application development.

The first application was developed in collaboration with the project team of the Little Bear River test bed in Utah (USU, 2008). The second and third applications are novel and were developed by the Iowa team through individual efforts and original solutions. The latter workflows can be used as blueprints or as starting-points in developing similar applications. The unique advantage of the applications developed for the digital observatories is that they can be easily shared, transferred and installed in servers of similar configuration (e.g. the Waters Network).

3.2. Real-time, automated extraction and data storage

One of the primary goals of the CCDW was to continuously ingest time series from heterogeneous data sources in a uniform format. This type of workflow was quite

Institution of Civil Engineers
IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

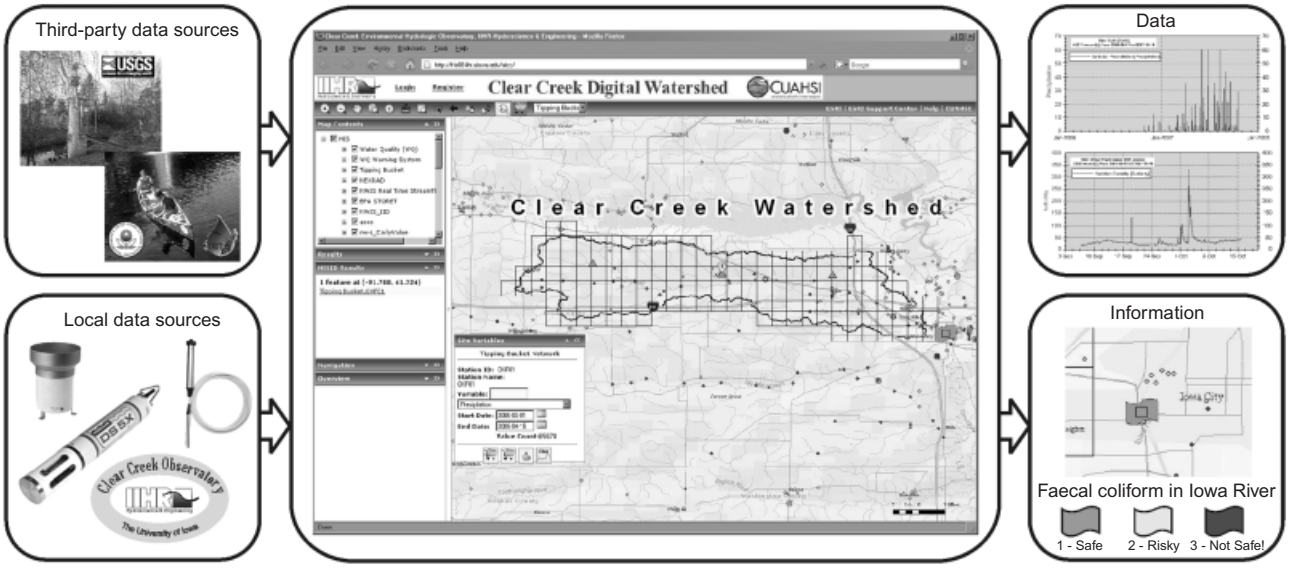

Figure 3. CCDW sample functions for interactive visualisation of data and information (UI, 2008b)

Data source	Variables*	Space scale		Time scale		Procedures/ comments
		Grain	Extent	Grain	Extent	
OTT Hydrolab Sonde 5X	N, P, DO, 3WQs	Point	n/a	20 min	08/31/2007–present	Local server
Tipping bucket	Precipitation	Point	n/a	15 min	08/15/2006–present	Local server
Climate	Precipitation, wind	1 km^2	270 km^2	1 h	01/01/2006–present	Local server (Hydro-Nexrad)
EPA Storet	Up to 11 WQs per site	Point	n/a	Sporadic	05/27/2000–present	www.epa.gov/storet
USGS NWIS	Q	Point	n/a	15 min	01/19/2000–present	river.sdsc.edu/wateroneflow
EPA-USGS NHDPlus	Hydrogeo datasets	Polyline	270 km^2	Static	Released in 2006	www.horizon-systems.com

*N, nitrogen; P, phosphorus; DO, dissolved oxygen; WQs, additional water quality variables; Q, stream flow

Table 1. Real-time data streams acquired and stored continuously in CCDW

straightforward to set up using Cuahsi-HIS tools for both third-party data (i.e. USGS, Environmental Protection Agency (EPA)) and observations acquired locally through the deployment of rain gauges (tipping bucket) and multi-task water-quality probes. The observations were stored in the Cuahsi-HIS ODM, which can be implemented in a variety of relational database management systems (RDBMSs). For the CCDW, Microsoft's SQL Server 2005 was used as the RDBMS in order to take advantage of the blank ODM template provided by the Cuahsi-HIS team. It should be noted that, although the ODM was designed to accommodate heterogeneous data into a single database, the choice was made to logically separate the data by observation network in order to minimise the interference that maintenance or corruption of one group of sensors could have on the entire database (e.g. all tipping bucket sites share one ODM database, which is separate from the database shared by the Hydro-Nexrad sites). In addition to the data model and the RDBMS, a tool was needed to ingest the raw observed data files into the ODM. For this purpose, the ODMDL and the SDL available in the Cuahsi-HIS software stack were used.

Loading significant amounts of data into a database by hand is tedious, error prone and not suitable for real-time applications. Several possible approaches were thus explored. The first attempt at populating the ODM database with observational data was to write custom software using the Java programming language. The customised software operates in two steps. First, an instrument data file was opened and parsed by the program in order to store timestamps and observational values into memory. Then, once the data were parsed, the program would proceed to make repeated SQL Insert statements to load the data value and its associated metadata into the database. Given the extent of the ODM structure, the amount of stored metadata was limited to the bare minimum in order to simplify the necessary programming. This, however, resulted in incomplete use of the ODM.

Following the early stages of the DW development, the group became aware of a more robust and complete alternative called ODMDL (To *et al.*, 2007). This software was designed to facilitate observational data loading into the ODM database via

ICE Institution of Civil Engineers
IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

a graphical interface. Although this software allowed for graphical loading of data into the ODM, it required that the data file be in a very specific MyDB format in order to be recognised. Therefore, another custom software package was developed to satisfy the constraints imposed by the ODMDL. This software was customised for specific data sources (e.g. tipping buckets) to open and parse the raw data files, transform the raw data into MyDB format and save the formatted data to a text file compatible with the ODMDL. This allowed the researchers to easily take raw data files and ingest them into the ODM database, with all the metadata and table relations set properly by the ODMDL software. A drawback of this approach is that it is still cumbersome to have to run the transformation program on each data file and manually load each of the resulting files into the ODM one by one using ODMDL.

While the ODMDL was usable for sporadic data observations (e.g. EPA water quality) with the detailed user's configuration, an additional tool was needed that was capable of taking data files from an instrument to the ODM database without interaction by the data manager in order to realise the goal of real-time data loading in the CCDW. For this purpose, Cuahsi-HIS developed the ODM SDL, which expanded the functionality of the ODMDL to include real-time operation and remove the need for the transformation to MyDB format prior to data loading.

The SDL software comprises two executable files – one with a graphical interface to allow for initial configuration and another that reads configuration information and processes data loading tasks. The SDL graphical configuration wizard makes it possible for a data manager to create a seamless, real-time link between the instrument's data file and the ODM database. The link is accomplished by mapping columns in the source data file to variable names in the ODM database. The mapping allows the SDL software to automatically load the values from the data file directly into the ODM without user interaction. It is necessary to choose a time frequency for each data file so the software will know how often it should check for new values. Once the mappings and frequency information are set, an XML configuration file containing those details is written to the local hard disk and subsequently used by the SDL executable. At this point, a way was needed to set the SDL executable to be run by the operating system on a regular basis. To accomplish this, Windows Task Scheduler was used to execute the SDL program every 15 min, which is the highest time frequency of the incoming real-time data files. The steps involved in populating the ODM with precipitation data measured by tipping buckets in the Clear Creek catchment are schematically presented in Figure 4.

3.3. Representation of spatially gridded data

Spatially gridded datasets such as GIS information, weather/climate grids and remote sensing are a common class of hydrologic data produced by instruments surveying large geographic areas. This type of data is not natively accommodated by the ODM, and therefore surrogate solutions that subvert this limitation were sought. Taking advantage of the Hydro-Nexrad project (Nexrad, 2008) carried out at UI, the team chose to represent precipitation estimates from Nexrad measurements with a resolution of 1 km^2. In order to maintain the usefulness and accessibility of the observations data, a server-based platform built on proven technology was required. The team chose the ESRI ArcGIS Server 9·2 platform (ESRI, 2008), which has built-in generalised web mapping application templates that can be customised to fit these needs. Extending these base templates can, however, require considerable effort and knowledge of several different technologies (JavaScript, Ajax, ASP, .Net, C#).

Instead of starting from scratch with an ESRI template, the group chose to leverage the Dash system developed by the Cuahsi-HIS project for the ArcGIS Server. This system comes with hydro-oriented data layers pre-installed, and provides data download and simple graphing capabilities. However, several challenges had to be overcome. The first was to find a method for taking data from the ODM database and making them available visually through Dash. While it is quite straightforward to represent a point observation in terms of its latitude and longitude, the problem becomes more challenging when spatial data, such as Nexrad data, need to be represented. Another difficulty involved storing spatial data into the ODM, because the data model was originally designed around the

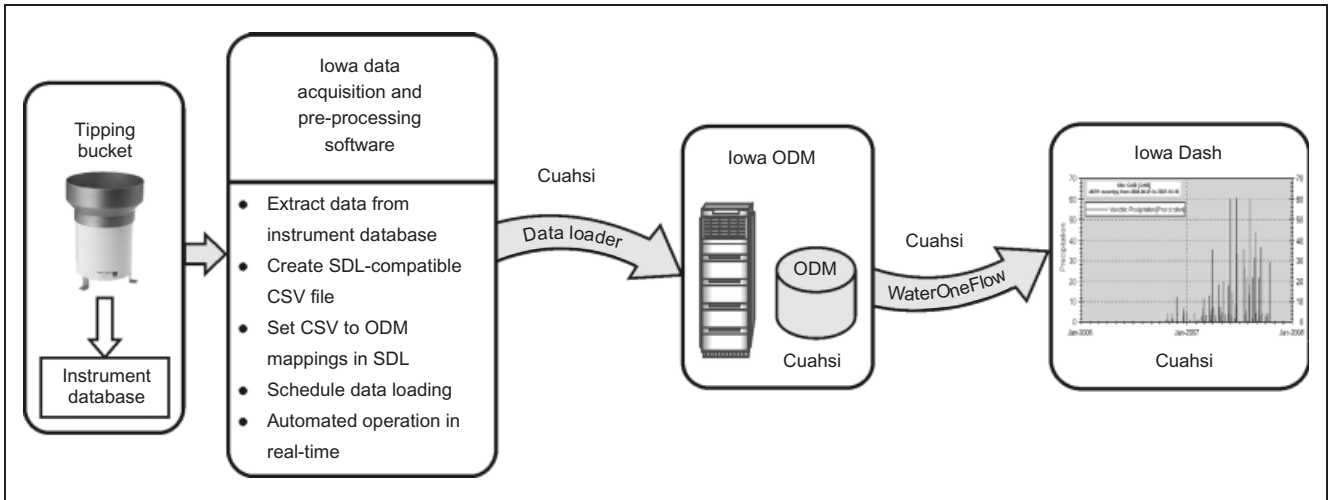

Figure 4. Flow diagram of data flux from sensor to representation of measurements with the tipping buckets deployed in Clear Creek

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

idea of storing time series data for point sites such as a gauging station or an instrument in a body of water.

The first obstacle of representing spatial Nexrad data in Dash was overcome by using ESRI's polygon feature type to represent each radar cell as a square (see Figure 5). Every centroid point inside the area enclosed by this polygon will contain the same data value. The size of these polygons is determined by the resolution of the radar data that they represent. In this case, each cell is 1 km^2 and 142 cells cover the study area. The next step was to figure out how to store Nexrad data into the ODM database. Since the grid of individual cells (represented as separate polygon features in Dash) was ready, each cell was considered as a 'site' in order to fit the data into the database. The latitude and longitude chosen for each polygon is that polygon's centroid. Utilising the centroid allowed the team to fit the spatial data into the ODM as a collection of points just like any other observation point.

Although there was now unique location information for each of the cells, it was also necessary to label each polygon with a site name and site code; a scheme that considered where the radar data came from was selected as appropriate. The source Nexrad data files come in ArcASCII format, which is a tab-separated text file with numerical data in rows and columns. Each [row, column] pair represents the data value for a particular cell in the larger grid that includes the area covering the Clear Creek catchment. For instance, location [44, 07] in the ArcASCII radar file contains the data value for the leftmost polygon that covers Clear Creek. From the original data file, this point is 44 cells to the right and seven cells down from the top-left origin. Since each cell in the grid will have a unique [row, column] pair, each site was named by concatenating the two values to form a site name such as 4407. The cell to the immediate right of 4407 would then be 4507 and the cell immediately above it would be 4406. For the sake of simplicity, the site name and site code for each cell were identical.

Using these storage and naming conventions allowed fitting spatial radar data into the ODM as a collection of 'points'. The benefit of this approach is that it enables leveraging all the Cuahsi community tools that rely on the ODM as their data source. If the team had chosen to store the Nexrad data in any alternative (geo) database, the Cuahsi software (Dash, WaterOneFlow, SDL, ODMDL, ODM Tools) could not be used. The successful gridded data representation using the Cuahsi-HIS components motivated new workflows for this type of data, such as dynamic visualisation and geotemporal queries on rainfall data layers. It was noted that, while getting a graph of rainfall estimation for a certain Nexrad bin can be useful, it is also sometimes desirable to be able to immediately see the

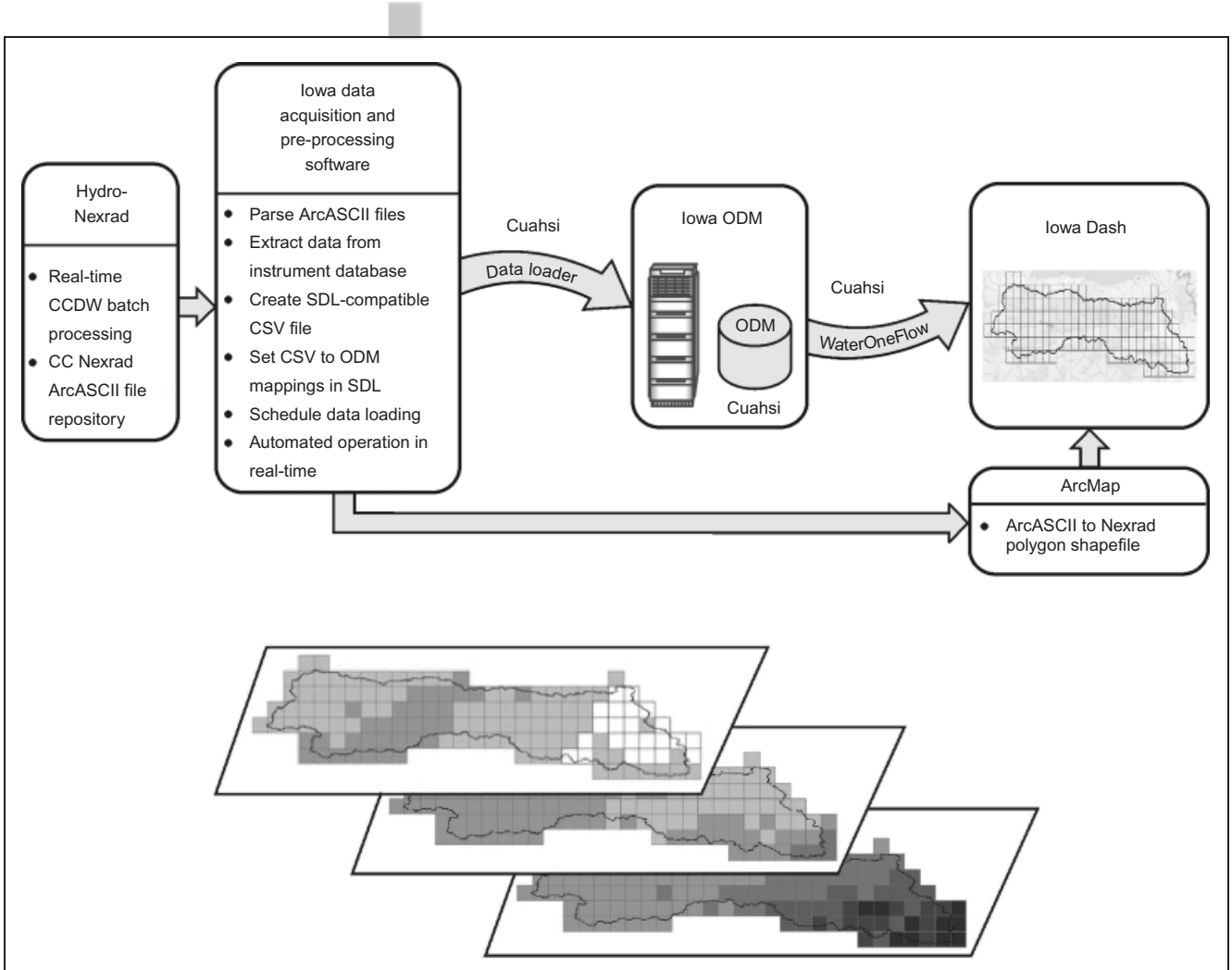

Figure 5. Flow diagram of data flux from Nexrad database to Cuahsi's Dash and visualisation of the radar cell in Dash

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

rainfall distribution for the entire study area. A new real-time module that updates the Nexrad shapefile with new hourly rainfall values was thus included. To accomplish this, the symbology of the map document was set to assign a colour code to each polygon based on the rainfall intensity for that area, as illustrated in Figure 5. Figure 5 also summarises the path of the Nexrad precipitation estimates from Hydro-Nexrad data source to their representation in the local server. The research team is currently working on developing geotemporal queries that will enhance the inference of useful information on rainfall dynamics over the area, as well as visual and numerical comparisons between rainfall estimations acquired with different instruments connected to the data model (i.e. tipping buckets and Nexrad).

While ODM and SDL worked successfully for the previous workflows, shortcomings were encountered when loading the spatially distributed Nexrad data. Such problems are symptomatic when workflows are implemented to larger catchments or areas of study that could contain hundreds or thousands of instrument data files. As described above, one data file was used for each of the 142 Nexrad bins that cover the Clear Creek catchment, and each of those bins represented in the ODM by one site. Clearly, it would be extremely tedious to use the graphical interface of SDL to configure the mappings for each of these data files one by one. Since each data file differs only by its filename, almost all of SDL's XML configuration information will remain the same across the entire dataset. It is immediately apparent that a program could be written to automate the process of creating the XML that would be appended to the SDL configuration file. Using a program to automatically generate the XML eliminated the need for a person to manually create mappings for all 142 Nexrad data files. In fact, only one data file needed to be manually mapped in order to utilise the generated XML as a template for the remaining files. After the configuration, XML was generated for the Nexrad files and appended to the existing config.xml file. The data files were then ready to be parsed and loaded by the SDL executable. It is considered that the lack of this feature in the SDL program could prove to be a barrier to its suitability and adoption for research groups with a large number of data collection sites.

Another shortcoming experienced during the initial data loading runs with Nexrad files was a considerable time lag. On checking Windows Task Manager, it was discovered that the data loader was using almost 700 MB of system memory and causing CPU usage to plateau around 100%. This equated to a nearly 100-fold increase in the use of system memory. Additionally, the sustained CPU usage required by the program caused all other processes in the system to lag severely. The entire data loading run took approximately 10 min, which was notably slower than the mere seconds taken when only a few data files needed to be parsed and loaded. Program lag and high memory usage was also observed on subsequent runs of the data loader, which eliminated the possibility that it was simply due to the extra processing that might be required for the first run. This use of the SDL to parse and load nearly 150 data files was atypical of the other test beds' usage; this indicates the SDL's limited ability to scale up for larger study areas and/or larger numbers of data files. Although further study could be conducted to yield a more accurate picture of SDL's limitations, experience has already shown that some optimisations will be necessary for the program to scale effectively when using a large number of data files.

3.4. Data assimilation in numerical simulations and information representation

One of the main targets for assembling the CCDW was to connect measurement data with numerical simulations to enable short- and long-term forecasting of water-related processes in catchments. For this purpose, the team selected a water-quality simulation model that, in conjunction with selected CCDW data streams, can predict *Escherichia coli* (E. coli) concentrations. Threshold values for concentrations were established to relate them to the level of acceptability for water quality for human health at the location of the measurement (see flags in Figure 3). Connecting observational data stored inside an ODM database with a model requires a software component to connect to the database to make queries on the data. Instead of writing code for this function, the team chose to use the WaterOneFlow web services described in Section 2. Dash was used to display the modelling results by creating a custom symbol and data layer to store the modelled values. An ArcMap plug-in developed by Cuahsi (called GetSites) was used to help with the creation of the new data layer. The modelled results were also stored into an ODM database. Accomplishing this workflow required the development and implementation of two complementary tasks that are now described.

3.4.1. Model integration. Customising numerical simulation models to run inside Dash as a web application extension is not trivial and requires knowledge of several technologies, including .Net, ArcObjects, XML and JavaScript. Therefore, the model software was implemented as a standalone real-time program that runs outside the process space of the web application. The selected water-quality model PhillyRiverCast (Maimone *et al.*, 2007) uses real-time turbidity, flow and rainfall data to provide public service information on the estimated current level of E. coli concentrations in the river and advise on the acceptable types of recreation based on those conditions. The prediction model was customised for the Iowa River to comprise a real-time data collection and modelling program written in Java, data file mapping with SDL and visualisation of the model's output in Dash, as illustrated in Figure 6. There are three components to the modelling program.

(*a*) The first component is responsible for collecting appropriate variables for the model via the WaterOneFlow services attached to the various ODM databases. The data streams come from sensors installed by the team or other agencies and are harvested as described in Sections 3.2 and 3.3.
(*b*) The second component is responsible for running the actual PhillyRiverCast model on these data streams to produce an integer value {1, 2, 3} that corresponds to the predicted bacteria concentration at the river site during that particular cycle of the model.
(*c*) The third component of the program is the CSV text file output. This file contains two columns: one for the timestamp and one for the integer prediction value. After

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

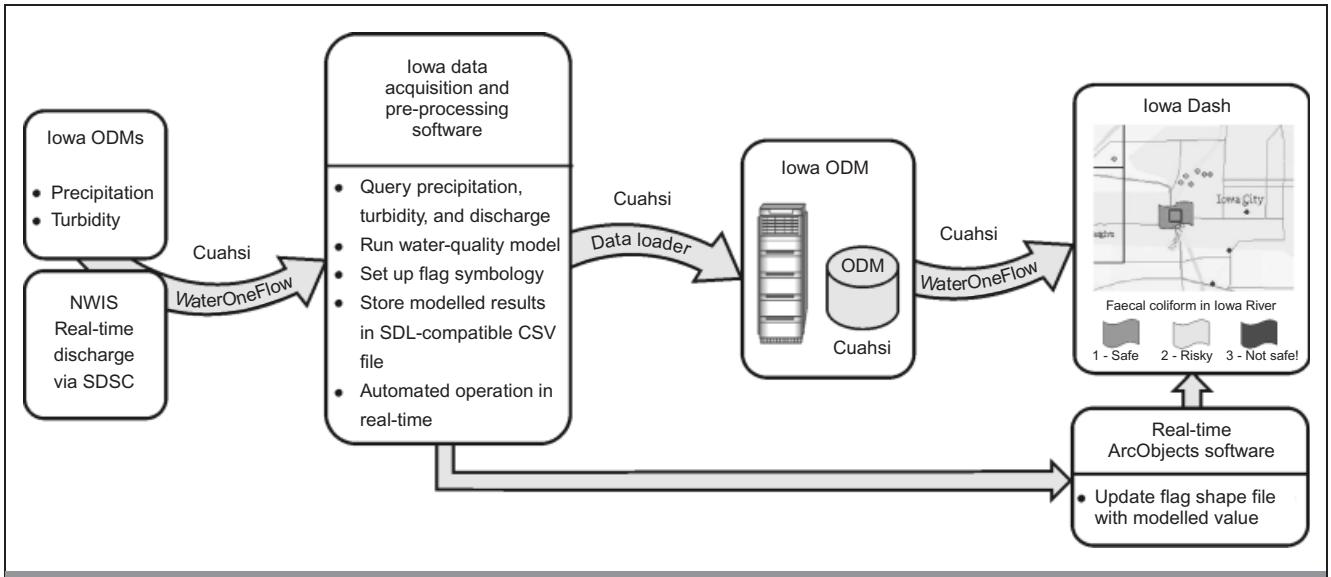

Figure 6. Illustration of the water-quality warning system workflow from model inputs to display of information regarding water safety level

appending the latest prediction value to the output file, the program sleeps for 1 h before repeating the process on new data.

3.4.2. Information representation. The intent of the warning system is to provide a straightforward graphical interface that can be easily understood by the public. The proposed graphic representation is a flag that changes colour hourly depending on the current water quality model prediction, as illustrated in Figure 6. The three possible flag colours are green, yellow and red, corresponding to model output threshold values considered safe, risky or unsafe. A green flag represents a situation where predicted E. coli levels are low enough to partake in activities involving water contact. A yellow flag indicates that water contact activities are not advised. Finally, a red flag indicates contact with water should be avoided.

Information regarding water quality is displayed using a graphic overlaid on the map interface inside Dash. The dynamic representation of the water quality flags in Dash was accomplished through a series of three steps. The first step involved using ArcMap to create a new feature in the Networks.mxd file. Use of the GetSites tool enabled the creation of a new 'site' on the map so the flag graphic could be displayed. The location for this feature has the same longitude and latitude as the collection site of the parameter data fed into the IowaRiverCast model. The second step entailed creating bitmap flag icons to correspond to the prediction values returned from the model. These bitmap files were then imported into ArcGIS with Styles Manager. In order for ArcGIS to display the correct flag symbol, the symbology of the newly created site had to be edited. This symbology setting created a link between the integer values that were stored in ODM and the three coloured flags that would represent this site visually. This link was created by adding a new field to the shapefile (e.g. RiverCastColor). This field will be populated with the most recent model prediction and will be used by ArcGIS to determine which colour flag to actually produce on the map. The third step – to get the colour of the flag to change based on the current prediction value – involved finding a way to update the RiverCastColor field inside the warning system's shapefile. To do this, a small program written in C# was developed that makes use of the ArcObjects software development kit (SDK) that is available as part of the ArcGIS software package. Once every hour the program queries the WaterOneFlow web service that is connected to the model's ODM database to extract the latest prediction value. Then the program utilises ArcObjects functions to open and update the RiverCastColor field in the shapefile previously mentioned. This final step is critical because without updating this field once per hour, the flag will not change colour in the map interface.

Although the current forecasting system works, there is room for improvement. Due to the nature of the numerical simulation, the process is not easily portable and would involve customisation if it were to be implemented in other research groups. A suggested improvement would be to streamline the entire process by creating one program that handles all the required functionality. This program would extract parameters for the model, run the model, output the CSV data file and update the shapefile once per hour. Because there is a steep learning curve associated with customising ESRI's web mapping applications, the team chose to implement the modelling software outside Dash as a real-time Java program. Efforts are currently being made to better understand how to extend and customise the Dash application, but this undertaking should not be attempted by groups without software development expertise.

4. DISCUSSION

The successful development of the workflows presented in this paper using an extensive number of components (but with limited involvement of the Cuahsi-HIS developing team) illustrates that the Cuahsi-HIS project has achieved its main role in supporting the water-related community in the management, publication and analysis of their data. The software package created by this community project removes the burden of learning and interpreting diverse file formats and then assembling them for further use from the data end user. The level of technical expertise and the amount of effort

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

required to develop customised workflows from scratch is not usually within the capabilities and resources typically available to local investigators. Nevertheless, the experience of the Iowa research group illustrates that the training materials written by the developers provide sufficient instruction to guide a user with limited computer skills to implement basic workflows such as loading data in an ODM, the use of Dash for visualisation and data downloading. However, customised workflows require computer proficiency and a considerable amount of effort.

Another positive experience for the local observatory team was the openness, readiness for collaboration and receptivity of the Cuahsi-HIS project team for user input. It was obvious from the project's inception that the information system was designed and built promoting a centralised development approach with broad community participation. This top-down developmental approach has provided the community with a uniform framework for data storage (the ODM) and common tools for data loading (ODMDL, SDL), data search, access and visualisation (Dash, HydroSeek) that otherwise would have been very difficult to build through individual (local) efforts. The easy and systematic access to data facilitates and the use of advanced analysis and synthesis tools enables users to focus more on the science of their work and spend less time simply trying to get their data in a usable or shareable form. The new tools were developed in parallel with user feedback and training activities (e.g. video seminars, workshops, email groups). The continuous two-way communication between the developers and users has facilitated the implementation of a complex information system into local observatories. Moreover, the formation of the data managers' community subgroup, with its own communication venues, has lessened the learning curve for local investigators engaged in this effort.

Clearly, the development of such a complex information system is a long-term process. As such, it is too early for this research team to make a sound cost–benefit assessment of using Cuahsi-HIS tools and functions for expanding the knowledge base instead of the traditional methods of handling data. Based on the experience gained from implementation and customisation of the CCDW system, the system is highly recommended to other local users, individuals or investigator teams. Initial developments have clarified the conceptual framework, highlighted practical benefits and identified future development needs.

Using the Cuahsi-HIS software stack in a local observatory

(*a*) enables assemblage, storage and publication of water observations acquired in local observatories in a standard way and federates them with third-party data providers such as specialised federal and state water-related agencies
(*b*) facilitates user access to more and better data for testing hypotheses and analysing processes from any place at any time over the internet within the user-preferred environment without the need to learn new software
(*c*) facilitates the real-time acquisition and storage of data
(*d*) relies on comprehensive and water-centric metadata embedded in the tool operation and specifications
(*e*) takes advantage of an elaborated syntactic and semantic mediation achieved through an extensive collaboration with the community and a carefully administrated ontology
(*f*) enables data discovery and sharing across observatories using uniform technology, standards and protocols
(*g*) enables knowledge and technology transfer within the water community through networking; the transfer can be made horizontally – among the servers – as well as vertically – between local servers and central HIS.

While the Cuahsi-HIS information system is still nascent and in full development, it is not yet widely spread in the community because

(*a*) initial efforts required to adopt and implement the system are not within the typical area of expertise of the hydrologic community
(*b*) there is inertia to change routine workflows without a demonstrated cost–benefit assessment of new ones
(*c*) there is potential conflict between generic and local data handling needs
(*d*) there is reluctance to abandon customised in-house developed workflows with a more rigid, community-geared information system that requires a long time for adjustment to local conditions.

In addition to coping with the above adverse conditions, the Cuahsi-HIS project team is addressing inherent system shortcomings in the early development stage. Some of the current Cuahsi-HIS component shortcomings have been discussed in recent data manager meetings (Waters Network, 2008b) and made into project priorities for the short-term research agenda (Maidment, 2008). The top research items are extensions of the ODM functionality and ancillary tools for the representation of spatially distributed data (which are ubiquitous in most of the hydrologic processes) and the addition of more advanced GIS-based visualisation and analysis capabilities. Improvements are also planned for simultaneously handling large amounts of data and the performance and functionality of the user interfaces. The community perception and feedback regarding the Cuahsi-HIS project is continuously incorporated into new versions of the training manuals in order to better address the research needs of the local observatories. It is expected that the dialogue between users and the central HIS team will continue, while the experience sharing between local data managers will expand commensurate with the rate of adoption of the new information technology.

5. CONCLUSIONS

This paper illustrates the path followed by the Iowa team in developing useful workflows that will hopefully be valuable to other groups implementing Cuahsi-HIS. While each of the applications presented here was developed for very specific purposes, the customisation process is quite generic and readily adaptable to conditions (i.e. sensor or sensor networks and models) in other observatories. The challenges faced by the research team are no different from those encountered by the other Waters Network test beds. The authors' experience was that it required expertise in computer science and an understanding of the Cuahsi-HIS technologies to build customised applications that are not covered in the training

Institution of Civil Engineers
IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

manuals (customising the Dash interface, making spatial data fit into ODM/Dash, etc.).

The successful deployment of the 11 Waters Network local observatories using Cuahsi-HIS products has shown that water observation data collected by academic investigators could be stored, published on the internet, federated with water observations data published by water agencies and searched using a concept framework that connects with variables in each individual data source. For many within the water resources community, the Cuahsi-HIS community project represents a new opportunity to approach the management, publication and analysis of their data systematically – that is, moving from collections of ASCII text or spreadsheet files to relational data models. While still searching for the best technological advancement in information science, the environmental engineering and hydrologic communities participate in activities aimed at enhancing the understanding and familiarity with emerging CI technology. Illustrative in this respect is the Cuahsi–Waters Network joint efforts to develop standardised data formats and a working service infrastructure, test various aspects of the design and operation of local/regional observatories and, eventually, create a national-scale network of observatories. It is expected that the type of synergy between the local observatories, water agencies and the scientific community projects illustrated in this paper will be critical in building the national observatory network.

Current and new community projects and local networks of observatories will continue to leverage the most recent advances in cyber-enabling technologies for assembling on-line data repositories, communication networks, advanced analytical and modelling tools, and powerful computing engines into a single digital environment for integrative research in water-related investigations. These cyber technologies facilitate integration approaches from hydroscience, engineering and water resource management, and one could suggest that the future development of these disciplines depends on this integration. The cyber platform could be gradually expanded to include other water resources allied disciplines (e.g. economics, geography and political science) to better capture the complexity of the dynamics of the multi-faceted water–human interaction. These end-to-end systems are prone to create a new paradigm for catchment science and management, enabling interdisciplinary teams to collaboratively understand and explain complex catchment issues.

ACKNOWLEDGEMENTS

The work presented in the paper was supported by the National Science Foundation project CBET 0607262. This support is gratefully acknowledged. Also acknowledged are the valuable comments of the anonymous reviewers that significantly improved this paper (reviewer 4 was especially thorough and helpful).

REFERENCES

Abbott MB (1991) *Hydroinformatics: Information Technology and the Aquatic Environment*. Avebury Technical, Aldershot.

Cuashi-HIS (Consortium of Universities for the Advancement of Hydrologic Science, Inc., Hydrologic Information System) (2008) See http://his.cuahsi.org/publications.html for further details (accessed 25/08/2008).

ESRI (2008) *GIS and Mapping Software*. See http://www.esri.com for further details (accessed 02/05/2008).

GIS WRC (GIS Water Resources Consortium) (2008) *ArcGIS Hydro Data Model*. See http://www.crwr.utexas.edu/giswr/hydro for further details (accessed 02/05/2008).

Goodall JL, Horsburgh JS, Whiteaker TL, Maidment DR and Zaslavsky IA (2008) A first approach to web services for the national water information system. *Environmental Modelling and Software* **23(4)**; 404–411.

Hall JW (2003) Handling uncertainty in the hydroinformatic process. *Hydroinformatics* **5(4)**; 215–232.

Hall JW and Anderson MG (2002) Handling uncertainty in extreme or unrepeatable hydrological processes – the need for an alternative paradigm. *Hydrological Processes* **16(9)**; 1867–1870.

Horsburgh JS (2007) *ODM Tools (Version 1·0)*. Environmental Management Research Group, Utah State University, Logan, UT, CUAHSI-HIS Document 3.

Horsburgh JS, Tarboton DG, Maidment DR and Zaslavsky I (2008) A relational model for environmental and water resources data. *Water Resources Research* **44**, W05406, doi:10.1029/2007WR006392.

Just C, Kruger A and Muste M (2007) The Clear Creek environmental hydrologic observatory: from vision toward reality. *AGU Fall Meeting*.

Liu J, Dietz T, Carpenter SR *et al*. (2007) Complexity of coupled human and natural systems. *Science* **317(5844)**; 1513–1516.

Maidment D (ed.) (2008) *Cuahsi Hydrologic Information System. Overview of Version 1·1*. Center for Research in Water Resources, University of Texas, Austin, TX.

Maidment D, To E, Smith V and Whiteaker T (2007) *Using the HIS Server Interface Dash (Version 1·0)*. Center for Research in Water Resources, University of Texas, Austin, TX, Cuahsi-HIS Document 7.

Maimone M, Crockett CS and Cesanek E (2007) PhillyRiverCast: a real-time bacteria forecasting model and web application for the Schuylkill River. *Journal of Water Resources Planning and Management* **133(6)**; 542–549.

Muste M (2007) Toward the integration of watershed science and management. *Proceedings* of the *32nd IAHR Congress, Venice* **2**, p522.

Muste M, Bennett D, Lawrence R and Kim D (2006) The Clear Creek environmental hydrologic observatory: from vision toward reality. *Proceedings of the 7th International Conference on HydroScience & Engineering, Philadelphia* 2006. See http://hdl.handle.net/1860/1420 for further details (accessed 02/12/2009).

Nexrad (2008) *Nexrad Rainfall Data for Hydrology*. See http://hydro-nexrad.net for further details (accessed 02/05/2008).

NSF (National Science Foundation) (2008) *Report of Blue-Ribbon Advisory Panel on Cyberinfrastructure*. See http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203 for further details (accessed 02/05/2008).

Price RK (2000) Hydroinformatics for river flood management. In *Flood Issues in Contemporary Water Management* (Marsalek J, Watt WE, Zeman E, Sieker F (eds)). Kluwer, Dordrecht, pp. 237–250.

To ESC, Whiteaker T and Valentine D (2007) *Loading Observations Data with the ODDataloader*. Center for Research in Water Resources, University of Texas, Austin, TX, Cuahsi-HIS Document 2.

UI (University of Iowa) (2008a) *Interdisciplinary Group for*

Institution of Civil Engineers
IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36

Cyberinfrastructure Development and Implementation. See http://www.iihr.uiowa.edu/CyberEnviroNet for further details (accessed 02/05/2008).
UI (University of Iowa) (2008b) *Clear Creek Digital Watershed*. See http://his08.iihr.uiowa.edu/uicc for further details (accessed 02/05/2008).
USU (Utah State University) (2008) *Little Bear River, Waters Test Bed*. See http://water.usu.edu/littlebearriver/for further details (accessed 24/08/2008).
Valentine D, Whitenack T, Whiteaker T and To E (2008) *Configuring Web Services for an Observations Database (Version 1·0)*. San Diego Supercomputer Center, University of California, San Diego, CA, Cuahsi-HIS Document 4.
Waters Network (2008a) *Science, Education, and Design Strategy*. See http://watersnet.org for further details (accessed 03/27/2008).
Waters Network (2008b) *Waters Testbed Data Managers*. See http://groups.google.com/group/waters-testbed-data-managers for further details (accessed 02/05/2008).
Whiteaker T (2007) *CUAHSI WaterOneFlow Workbook (Version 1·0)*. Center for Research in Water Resources, University of Texas, Austin, Cuahsi-HIS Document 5.
Whitenack T (2007) *Getting Started with the HIS Server (Version 1·0)*. San Diego Supercomputer Center, University of California, San Diego, CA, Cuahsi-HIS Document 1.

What do you think?
To discuss this paper, please email up to 500 words to the editor at journals@ice.org.uk. Your contribution will be forwarded to the author(s) for a reply and, if considered appropriate by the editorial panel, will be published as discussion in a future issue of the journal.

Proceedings journals rely entirely on contributions sent in by civil engineering professionals, academics and students. Papers should be 2000–5000 words long (briefing papers should be 1000–2000 words long), with adequate illustrations and references. You can submit your paper online via www.icevirtuallibrary.com/content/journals, where you will also find detailed author guidelines.

IP: 128.255.53.136
Fri, 22 Jan 2010 21:03:36