International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Hydrologic Data Discovery, Download, Visualization, and Analysis: A Brief Introduction to HydroDesktop

Daniel P. Ames[a], Jeffery S. Horsburgh[b], Yang Cao[a], Jiri Kadlec[a], Timothy Whiteaker[c], David Valentine[d]
[a]*Ctr. for Adv. Energy Studies, Idaho State Univ., Idaho Falls, Idaho, USA.*
[b]*Utah Water Research Laboratory, Utah State Univ., Logan, Utah, USA*
[c]*Center for Research in Water Resources, Univ. of Texas at Austin, Texas, USA*
[d]*San Diego Supercomputer Center, University of California, San Diego, USA.*
[*]*Corresponding Author Email: dan.ames@isu.edu, +1-208-533-8141*

Abstract: Discovering and accessing hydrologic and climate data for use in research or water management can be a difficult task that consumes valuable time and personnel resources. New advances in cyberinfrastructure and in semantic mediation technologies have provided the means for creating better tools supporting data discovery and access. In this paper we describe a freely available and open source software tool, called HydroDesktop that can be used for discovering, downloading, managing, visualizing, and analyzing hydrologic data. HydroDesktop was created to discover data published using the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS).

***Keywords**:* data management, geographic information system, hydrologic information systems, hydrologic modeling, observation data, web services.

1 INTRODUCTION

In this paper, we briefly introduce the design, architecture, and implementation of a newly developed open source Hydrologic Information System (HIS) software tool called HydroDesktop. HydroDesktop was designed to enable the discovery and retrieval of syntactically homogenous data hosted on any of the distributed hydrologic data servers registered with the Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI) HIS system using user-specified spatial, temporal, and keyword based constraints to narrow search results. Through an extensible graphical user interface (GUI), HydroDesktop provides many capabilities needed by hydrologic data consumers, including: discovery of hydrologic time series data; map-based visualization of monitoring locations and other geographic information systems (GIS) data; download, organization, visualization, editing, and maintenance of hydrologic time series; linkage with integrated modeling systems (OpenMI); and linkage with common data analysis and modeling software such as the R statistical computing environment.

2 BACKGROUND

The challenges of discovering and integrating disparate data and schemas from physically distributed sources are not unique. Finding solutions to interoperability problems is a common component of large cyberinfrastructure projects being conducted within many scientific domains, including geology [Nambiar et al. 2006], oceanography [Chave et al. 2009], and atmospheric sciences [Droegemeier et al. 2004]. Within the hydrology domain, there have been no public standards for data organization, formats, or publication mechanism that would increase interoperability

of water observations data expressed as time series. Consequently, there has been no means of unified discovery or access to water observations information.

A central challenge in seamlessly integrating multiple data sources is resolving heterogeneity issues [Beran and Piasecki 2009; Piasecki and Beran 2009; Horsburgh et al. 2009]. Unique data sources may use different vocabularies to describe data collection locations or measured variables, making it difficult to search multiple systems for similar data. This can be difficult when a study requires data from multiple scientific domains, and the scientist is not familiar with the vocabulary used by one or more of these domains. Performance of queries and search mechanisms for data discovery can be significantly improved when semantic heterogeneity in data among datasets is overcome [Madin et al. 2007].

Beran and Piasecki [2009] described several innovations within this problem space, and indeed, much of the work related to discovery of hydrologic data described in this paper is an outgrowth of their work. They described an ontology-aided search engine website called HydroSeek, which was developed to enable users to query multiple hydrologic data repositories simultaneously using keywords. They created this functionality by developing a "knowledge base" that covered the water quality, meteorology, and hydrology domains, and providing a linkage between scientific or everyday language (e.g., the keywords or terms that scientists would use to search for data) and the language and variable codes used by repositories to store data. The data discovery and download functionality of HydroSeek has been generalized and expanded within the CUAHSI HIS system through the creation of two major components. The first is a central web service registry, metadata catalog, and data discovery service called HIS Central. The second is HydroDesktop, which is end-user client software that has data discovery and download capabilities similar to HydroSeek as well as a number of data organization, management, analysis, visualization, and management tools.

3 A SERVICE-ORIENTED ARCHITECTURE FOR HYDROLOGIC OBSERVATIONS

The design of the CUAHSI HIS (Figure 1) follows an open, service-oriented architecture (SOA) model. SOA relies on a collection of loosely-coupled, self-contained services that communicate with each other through the Internet and can be called from multiple clients in a standard fashion [Goodall et al. 2008]. Common benefits associated with a SOA include: scalability, security, easier monitoring and auditing, standards-reliance, interoperability across a range of resources, and plug-and-play interfaces [Josuttis 2007, Goodall et al. 2010]. Internal service complexity is hidden from service clients, and backend processing is decoupled from client applications, making the core of the system independent of a specific platform or implementation [Huhns and Singh 2005; Granell et al. 2009]. As a result, different client applications are able to access the same service functionality, leading to a more modular, transparent, and easier to manage system.

The CUAHSI HIS infrastructure represents a collection of computer servers, referred to as HydroServers, which support publishing hydrologic observations data [Horsburgh et al. 2010]. Over the past several years, HydroServers have been installed at various universities and public agencies, and there are now a large and growing number of hydrologic observations data available via web services published on HydroServers [Horsburgh et al. 2009]. Each HydroServer exposes one or more web services, called WaterOneFlow, that provide uniform access to multiple repositories of water observations data. Each web service contains two types of methods: 1) a data delivery method called GetValues, which publishes the values of water observations; and 2) metadata delivery methods, including GetSites, GetSiteInfo, and GetVariableInfo, which identify and describe collections or series of data values associated with particular spatial locations. WaterOneFlow web services publish data on a HydroServer using a markup language called Water Markup Language (WaterML) [Zaslavsky et al. 2007].

Data published on HydroServers using WaterOneFlow web services are indexed within a central web service registry and metadata catalog called HIS Central. HIS Central regularly harvests the metadata describing published time series of hydrologic observations from registered WaterOneFlow web services by calling their metadata delivery methods. The metadata are compiled within a central metadata catalog database, which is then exposed to search queries via a data discovery web service. HIS Central also contains a variable name ontology that is used to tag variable metadata harvested from WaterOneFlow web services and enable mediation across the vocabularies used by different data sources. This data discovery web service is publicly accessible and can be called by any client application that wishes to incorporate data discovery capabilities. It supports spatial, temporal, and variable keyword constraints to narrow search results.

The CUAHSI HIS SOA is completed by client applications that use the data discovery service available at HIS Central to enable discovery of relevant time series of hydrologic observations and then use the metadata returned by data discovery queries to access the data via the WaterOneFlow web services on the HydroServer that hosts the data.

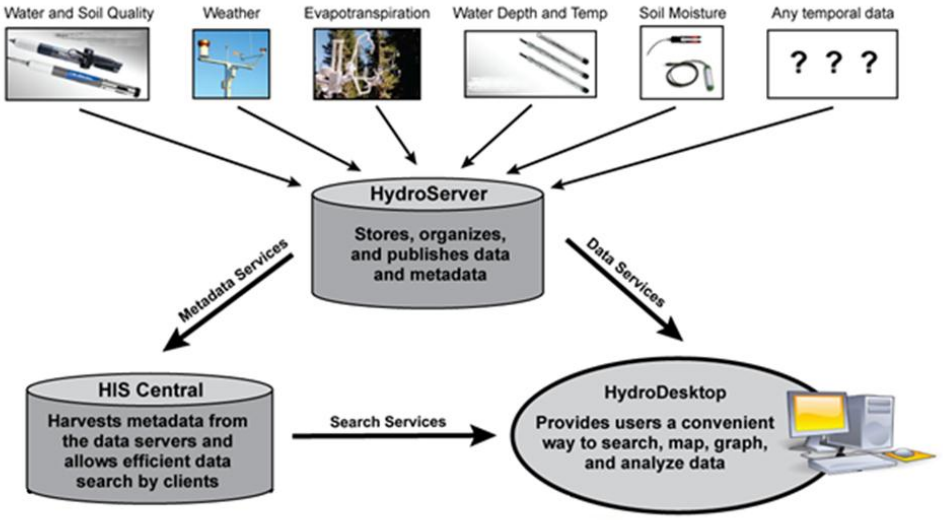

Figure 1. Key components of the CUAHSI HIS. HydroServer contains a database of observations, a web service for sharing observations data. There are multiple Hydroservers. HIS Central is a registry of publicly accessible HydroServers that provides data discovery and search web services. HydroDesktop is a client application with a GIS-based user interface for data discovery, retrieval and analysis (Figure courtesy of Stephen Brown, Univ. of New Mexico.)

4 ARCHITECTURE, DESIGN, AND KEY CAPABILITIES OF HYDRODESKTOP

HydroDesktop serves as a common window into observational data published using WaterOneFlow web services [Ames et al. 2009]. Data discovery is accomplished through searches across a comprehensive metadata catalog maintained at HIS Central or individual HydroServers hosting WaterOneFlow web services. These searches are facilitated by additional web services that expose the metadata catalog and the Hydrologic Ontology maintained at HIS Central. Search results can be refined to specify datasets that a user would like to download. Data downloads are performed by making GetValues calls to the appropriate WaterOneFlow web services. Downloaded data are stored in a desktop data repository database following a relational database schema which is accessible to

additional tools and software through an application programming interface (API) or directly.

Visualization and analysis tools that are part of HydroDesktop are developed using the API data access method to maintain a level of data access consistency and integrity as well as abstraction from the HydroDesktop database. Additionally, users can access data through third party data analysis applications that have the ability to read from a relational database (e.g. R and MATLAB). HydroDesktop includes plug-ins developed by the HIS team and also supports third party plug-ins that follow a well-defined plug-in interface described at the project web site.

HydroDesktop has a primary interface similar to most desktop GIS programs with the addition of tools and forms specifically related to time series data visualization and analysis. Included are a ribbon-style main menu, legend, and a main map display in a tabbed interface with movable/dockable panels (Figure 2). The map display is the main visualization element, while the other portions of the interface provide tools for searching, obtaining, and managing data. HydroDesktop was developed with a simple interface that should be easily usable regardless of the operator's technical background. The GIS capabilities are powered by the open source DotSpatial GIS components (see www.dotspatial.org). The primary purpose of HydroDesktop is to facilitate discovery and access of hydrologic data. A secondary purpose is to provide support for data manipulation and synthesis. The user interacts with HydroDesktop via a GUI with the functionality described below.

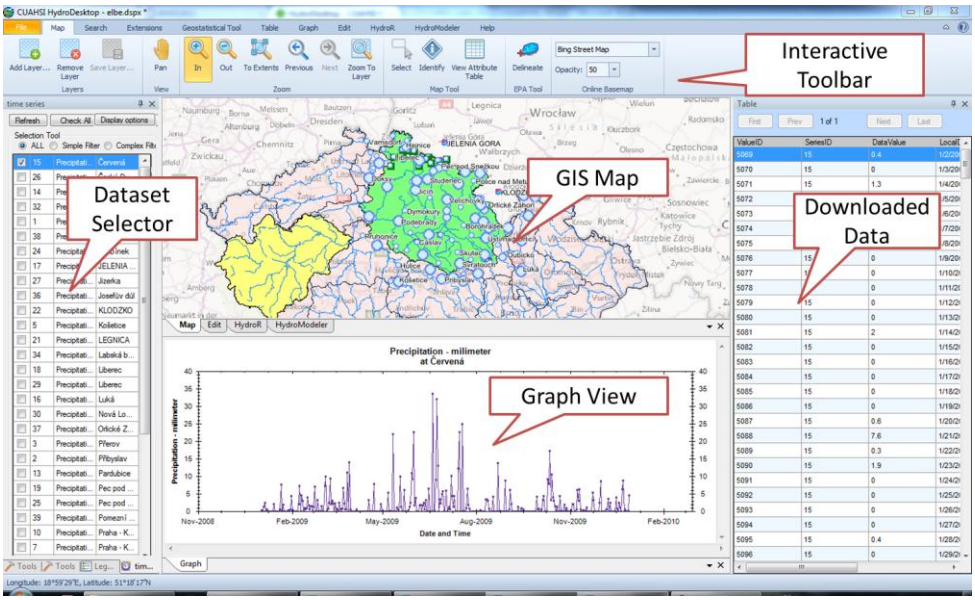

Figure 2. Extensible ribbon based layout design of HydroDesktop.

4.1 Data Discovery

HydroDesktop supports two different methods of data discovery: 1) ontology-based discovery across all WaterOneFlow web services that have been registered at HIS Central and for which metadata has been harvested and stored in the HIS Central metadata catalog; and 2) discovery of data within a single WaterOneFlow web service that has not been registered at HIS Central. The first type of data discovery is supported by HIS Central metadata web services that expose the contents of the HIS Central metadata catalog. The second type of data discovery involves making data discovery calls directly to the web service that has not been registered with HIS Central. This approach facilitates both the use of datasets cataloged and documented at HIS Central, as well as use of datasets stored on individual or regional HydroServers but not registered with HIS Central.

HIS Central includes a metadata catalog describing the time series datasets served by registered WaterOneFlow web services. This catalog includes the mappings between variables and HIS Ontology concepts. This catalog is automatically updated weekly and represents a comprehensive listing of data published using WaterOneFlow services and registered at HIS Central. The contents of the HIS Central metadata catalog are exposed by a web service API that provides methods for retrieving the following information: (i) the full metadata description (including the URL to the WaterOneFlow service) for all WaterOneFlow web services registered at HIS Central, (ii) a listing of all searchable keywords/concepts from the HIS Ontology, and (iii) the full metadata description for all data that meet certain spatial, temporal, and variable search criteria.

HydroDesktop uses the methods from the HIS Central metadata catalog API to search for data series that meet criteria input by a specific user. HydroDesktop presents users with a search tool that supports the following search criteria: (i) a latitude/longitude bounding box to serve as the spatial constraint on the query. The box can be input by typing in coordinates, by drawing a rectangle on the HydroDesktop map, or by selecting a polygon feature from one of the layers in the HydroDesktop map (e.g., a watershed boundary – the extent of the feature would be converted to a latitude/longitude box), (ii) a searchable concept from the HIS Ontology (to be input by the user or selected from a list), (iii) a begin date and end date to serve as the temporal constraint on the query, (iv) a minimum number of observations (only data series that have more than this minimum number for the entire data record will be selected, regardless of time window specified), and (v) a list of WaterOneFlow web services to include in the search. This will be a user-specified subset of the web services registered at HIS Central that constrains search results to only a selected set of web services.

The result of a data discovery query using the HIS Central metadata catalog is the full metadata description for a listing of all of the data series cataloged at HIS Central that meet the search criteria. For example, a user may choose to search the entire state of Utah for streamflow data. The results of the search will be a list of sites and data series that meet the criteria. The user can then subset the results to the data series of particular interest, i.e. after seeing a map of the locations of several hundred streamflow gauge sites in Utah, the user may choose to only retrieve data for sites that meet some additional condition. The user then organizes data into a thematic data set on the local machine for viewing and interaction.

4.2 Data Download

The goal of the HydroDesktop data download functionality is to retrieve observational data series that have been identified for download using the data discovery tools described above and to create a local cache copy of the data in the local database. The metadata resulting from the discovery process consist of a descriptive list of data series identified by a user for download. Using this list, HydroDesktop issues GetValues calls to retrieve each data series in WaterML format. HydroDesktop saves a copy of the result of each GetValues call as a WaterML formatted XML file on the user's hard drive. Next, HydroDesktop parses each of the WaterML results into the HydroDesktop data repository database.

The data repository database has a relational structure and is implemented within a relational database management system (RDBMS), serving as a local cache copy of the data that have been retrieved. The relational schema of the data repository database is semantically similar to the CUAHSI ODM database design [Horsburgh et al. 2008], with similar naming conventions and data types, but has been modified and extended to facilitate management of the data series that have been downloaded and storage of provenance information. The relational database schema of the HydroDesktop data repository database is available for viewing at http://hydrodesktop.codeplex.com/ documentation.

The data repository database is capable of storing all of the information encoded within WaterML files resulting from GetValues calls and also supports the storage of provenance information, including: (i) where was the data obtained, i.e., which web service? (ii) the query that resulted in the data that was loaded (the GetValues call used to get the data); (iii) a pointer to the WaterML file from which the data originated (the file is cached locally); (iv) the date on which the data were loaded; (v) the last date on which the data were checked for updates; (vi) the last date on which the data were updated with new data; and (vii) what has been done to the data since it was added to the database.

4.3 Data Visualization, Editing, and Export

HydroDesktop supports visualization of both geospatial and time series data through an interactive GIS map using the open source DotSpatial GIS components which are based on the MapWindow GIS system [Ames et al. 2008] and 3rd party DotSpatial plug-ins. DotSpatial supports a variety of vector, raster, and image GIS data types, and includes functionality for navigating the map as well as many other GIS tools and features. The HydroDesktop interactive map is used for displaying and manipulating spatial datasets as well as for setting the context for data discovery. As described in the sections above, an area of interest is often used as a spatial filter for narrowing a search for data. The interactive map enables the user to set the geographic context for data discovery and access by enabling users to draw a bounding box or select a polygon feature from one of the GIS layers in the map within which they would like to conduct their search.

Once time series of observational data have been retrieved and stored in the desktop data repository database, HydroDesktop provides users with tools for visualizing and analyzing the data. HydroDesktop maintains a GIS data layer showing the locations of the sites for which data have been downloaded to the desktop data repository database. This layer is dynamically built from the data repository database each time data are downloaded. Visualization of observational data is provided through a variety of plots using the open source Zed Graph plotting package and is focused on exploratory data analysis for data series that are downloaded and stored in the HydroDesktop data repository. Plot types available for visualizing time series data at a selected site include time series, histogram, box-and-whisker, and probability plots. The HydroDesktop time series visualization and analysis tool also enables users to view a selected time series in a simple tabular view and calculates simple descriptive statistics for the selected time series.

Additionally, HydroDesktop includes an R statistics plug-in that supports manipulation and transformation of data, statistical analysis, and modeling using data from the HydroDesktop database. A data export plug-in allows users to export selected observational data from the local database to a delimited text file. The HydroModeler plugin provides a graphical interface for constructing hydrological simulation model workflows from process components (such as infiltration, evapotranspiration) that can use selected observational data as input [Castronova et al. 2010].

5 DISCUSSION AND CONCLUSIONS

The main contributions of this work are: (i) HydroDesktop provides free access to data from distributed data services that are part of the CUAHSI HIS Internet-based, service oriented architecture (SOA) and its 23 million data series; (ii) the HydroDesktop software interface enables end users that include university faculty, graduate and undergraduate students, K-12 students, engineering and scientific consultants, and others to operate within a relatively uncomplicated software environment; (iii) as an open source, free software application, HydroDesktop does not require use of commercial, third party software beyond the operating system and hence is expected to facilitate growth of a community of users and developers

who can maintain and enhance the software. An on-going usability study focused on improving HydroDesktop and demonstrating/quantifying its efficiencies and performance over legacy methods is also underway, and results will be published. While the core HydroDesktop software is complete and available for use (over 33,000 downloads as of March 2012), new plug-ins and extended capabilities are under active development at http://hydrodesktop.codeplex.com/. Here project participants, both from the CUAHSI HIS team and volunteers from the hydrologic sciences community share a discussion forum, bug tracking system, documentation WIKI, and an open Mercurial code sharing repository. User support and documentation for HydroDesktop is provided informally by the open source and volunteer development community at the project web site (including step-by-step tutorials) as well as formally through a series of workshops, webinars, and outreach activities sponsored by CUAHSI (see http://his.cuahsi.org) and through the detailed help system included with the software.

Any interested parties are invited to visit the project website, download the source code and join in the development and testing activities related to this project. It is expected that the simple plug-in architecture will encourage and facilitate third party development of plug-ins that significantly extend the base HydroDesktop application, making full use of all of the data retrieval and storage mechanisms in the initial version of HydroDesktop. Specific future development plans for HydroDesktop include: support for new data sources and formats (including the OGC WaterML 2.0 standard); entry and upload of data into a HydroServer via HydroDesktop (e.g. for data collection purposes); ability to find and view metadata for datasets with limited access rights; a number of geospatial data analysis tools provided through the DotSpatial toolbox (e.g. geostatistical interpolation, clipping); and new time series management tools (e.g. unit conversion).

6 SOFTWARE AVAILABILITY

All CUAHSI HydroDesktop software and documentation can be accessed at http://his.cuahsi.org. Source code and additional documentation for HydroDesktop can be accessed at the HydroDesktop code repository website http://hydrodesktop.codeplex.com.

ACKNOWLEDGMENTS

The software described in this paper was developed as part of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) project. This work was supported by the National Science Foundation under grant EAR 0622374 and under the Idaho NSF EPSCoR grant EPS 0814387.

REFERENCES

Ames, D.P., Michaelis, C.D., Anselmo, A., Chen, L., Dunsford, H., 2008. MapWindow GIS, in: Shekhar, S., Xiong, H. (Eds.), *Encyclopedia of GIS*. Springer, New York, pp. 633-634.

Ames, D.P., Horsburgh, J.S., Goodall, J., Tarboton, D.G., Whiteaker, T., Maidment, D.R., 2009. Introducing the open source CUAHSI Hydrologic Information System desktop application (HIS Desktop). In Anderssen, R.S., R.D. Braddock and L.T.H. Newham (eds) *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation*, July 2009, pp. 4353-4359. ISBN: 978-0-9758400-7-8. http://www.mssanz.org.au/modsim09/J4/ames.pdf.

Beran, B., Piasecki, M., 2009. Engineering new paths to water data. *Computers and Geosciences* 35, 753-760, doi:10.1016/j.cageo.2008.02.017.

Castronova, A.M., and Goodall, J.L., A generic approach for developing process-level hydrologic modeling components, *Environmental Modelling & Software*, Volume 25, Issue 7, July 2010, Pages 819-825, ISSN 1364-8152, 10.1016/j.envsoft.2010.01.003.

Chave, A.D., Arrott, M., Farcas, C., Farcas, E., Krueger, I., Meisinger, M., Orcutt, J.A., Vernon, F. L., Peach, C., Schofield, O., Kleinert, J.E., 2009. Cyberinfrastructure for the US Ocean Observatories Initiative: Enabling interactive observatories in the ocean. In: *Proceedings of the OCEANS '09 IEEE Conference, Bremen, Germany*. IEEE Ocean Engineering Society, pp. 1–10,doi:10.1109/OCEANSE.2009.5278134.

Droegemeier, K.K., Chandrasekar, V., Clark, R., Gannon, D., Graves, S., Joseph, E., Ramamurthy, M., Wilhelmson, R., Brewster, K., Domenico, B., Leyton, T., Morris, V., Murray, D., Plale, B., Ramachandran, R., Reed, D., Rushing, J., Weber, D., Wilson, A., Xue, M., Yalda, S., 2004. Linked Environments for Atmospheric Discovery (LEAD): A cyberInfrastructure for mesoscale meteorology research and education. In: *Proc. of the 20th Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, American Meteorological Society, Seattle, Washington, Amer. Meteor. Soc., CD-ROM, S6.1.

Goodall, J.L., Horsburgh, J.S., Whiteaker, T.L., Maidment, D.R., Zaslavsky, I., 2008. A first approach to web services for the National Water Information System. *Environmental Modelling & Software* 23 (4), 404-411, doi:10.1016/j.envsoft.2007.01.005.

Goodall, J.L., Robinson, B.F., Castronova, A.M., 2010. Modeling water resource systems using a service-oriented computing paradigm. *Environmental Modelling & Software*, 26 (5), 573-582, doi:10.1016/j.envsoft.2010.11.013.

Granell, C., Diaz, L., Gould, M., 2009. Service-oriented applications for environmental models: Reusable geospatial services. Environmental Modelling & Software, 25 (2), 182-198, doi:10.1016/j.envsoft.2009.08.005.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. *Water Resources Research* 44, W05406, doi:10.1029/2007WR006392.

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., Whitenack, T., 2009. An integrated system for publishing environmental observations data. *Environmental Modelling & Software* 24 (8), 879-888, doi:10.1016/j.envsoft.2009.01.002.

Horsburgh, J.S., Tarboton, D.G., Schreuders, K.A.T., Maidment, D.R., Zaslavsky, I., Valentine, D., 2010. Hydroserver: A platform for publishing space-time hydrologic datasets. In: *Proceedings 2010 AWRA Spring Specialty Conference GIS and Water Resources* VI, Orlando, Florida, AWRA, Middleburg, Virginia, TPS-10-1.

Huhns, M., Singh, M., 2005. Service-Oriented Computing: Key Concepts and Principles. *IEEE Internet Computing* 9 (1), 75-81.

Josuttis, N. M., 2007, SOA in Practice: The Art of Distributed System Design, O'Reilly Press, Sebastapol, CA, 324 p.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. *An ontology for describing and synthesizing ecological observation data*. Ecological Informatics 2 (3), 279–296, doi:10.1016/j.ecoinf.2007.05.004.

Nambiar, U., Ludaescher, B., Lin, K., Baru, C., 2006. The GEON portal: Accelerating knowledge discovery in the geosciences. In: *WIDM '06 Proc. of the 8th Annual ACM Workshop on Web Information and Data Management*, Arlington, Virginia, USA. Association for Computing Machinery Press, pp. 83-90,

Piasecki, M., Beran B., 2009. A semantic annotation tool for hydrologic sciences. *Earth Science Informatics* 2 (3), 157-168, doi:10.1007/s12145-009-0031-x.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.), 2007. CUAHSI WaterML. OGC Discussion Paper OGC 07-041r1. Version 0.3.0. <http://portal.opengeospatial.org/files/?artifact_id=21743>. (16.11.10.).