**Requirements for Supporting Access Control for Published**

**Point Observations Data Using the CUAHSI HIS**

Jeffery S. Horsburgh, David G. Tarboton, Kimberly A. T. Schreuders, John Cullen

8/16/2010

# 1. Introduction

Academic researchers who are collecting data within experimental watersheds, observatories, and research sites have need for the ability to control access to both private and public data resources in their data collection, management, and publication process. One mechanism for publishing data from experimental sites is using a CUAHSI Hydrologic Information System (HIS) HydroServer. To address the need for controlling access to data, new functionality is needed that enables HydroServer administrators and data owners to control access to both public and private data resources hosted on a HydroServer.

The current model for publishing point observations data using HydroServer is to load the data into an Observations Data Model (ODM) database, implement the WaterOneFlow Web services, and register the services with HIS Central. Once the Web services have been registered with HIS Central, the entire contents of the ODM database are effectively published on the Internet, and there is no restriction on or who can access the data through HydroServer publication software such as the WaterOneFlow web services and Time Series Analyst. There are a number of situations where data producers want to take advantage of the organization and functionality that the current HydroServer Software Stack provides, but without providing unrestricted and unlogged access to all of the data resources that they are putting on their server. These include the desire of academic data collectors/publishers to:

1. Have better control over how, when, and if data go from private to public.
2. Publish research results (e.g., peer reviewed publications) based on data before the data are released to the general public.
3. Keep track of who is downloading and using their data to evaluate and document its impact on the community.
4. Have and use a data use/access agreement and ensure that they get credit and appropriate citation for the data that they publish.
5. Control who can access/download data.
6. Only expose the best or highest quality data, while restricting access to preliminary or raw versions of the data.
7. Integrate their data organization, management, and publication rather than maintaining separate systems for each of these functions.

The following sections provide definitions of the terminology used and details about the types of data that need to be supported and the functionality that is needed for access control.

# 2. Definitions

The following are specific definitions for some of the terminology in this document:

***Access Control*** – The overall process by which data owners/providers control which data consumers can access their data resources.

***Authentication*** – The process by which a data consumer provides credentials that verify his or her identity.

***Authorization*** – The process by which a data owner/provider enables a data consumer to access/download a data resource.

***Client Application*** – A software program that downloads data from a HydroServer.

***Data Consumer*** – A person that wants to download data hosted on a HydroServer.

***Data Owner*** – The person or organization that collected the data and wishes to control its distribution.

***Data Resource*** – An identifiable unit of data to which access control rules can be applied and that can be downloaded.

***HydroServer Administrator*** – The person or persons that have full administrative access to a HydroServer and that have authority to grant access control privileges.

***Registration*** – The process by which a data consumer creates a user account that can be given access to data resources.

# 3. Metadata and Classes of Data

For the purposes of this document, “metadata” is defined as all of the information stored in an ODM database except for the actual data values (e.g., the existence of data series that have been measured and information about their sites, variables, methods, sources, quality control levels, etc.). Currently, metadata about data series stored in an ODM database and published through the WaterOneFlow web services can be automatically harvested and cataloged at HIS Central.

In considering support for access control over datasets published on a HydroServer, the following types of data are anticipated:

1. Type 1 – Public Data No Restrictions: The metadata are public, the data are discoverable by the public, and the data are accessible/downloadable by the public with no data use agreement, no tracking, and no restrictions. No data consumer authentication is required. Identification of data consumers is not required, and is often not desirable.

Example 1: Data from public agencies like USGS NWIS or EPA STORET.

1. Type 2 – Tracked Public Data: The metadata are public, the data are discoverable by the public, and the data are accessible/downloadable by the public under a “general” data use agreement. A search will reveal all locations of all variables, sites, and data series. Authentication of data consumers would be required for downloading the data, but only so data access/use could be tracked. The general data use agreement would indicate the conditions of use for the data and serve to notify users of the tracking of downloads and how tracking information will be used fulfilling any privacy and information use notification requirements that may exist.

Example 1: An academic investigator makes data from publicly funded research available online under a general data use agreement. The investigator wants to track who is downloading and using the data to establish the broader impact of his/her datasets and ensure appropriate citation.

1. Type 3 – Public Metadata with Private Data – The metadata are public, and the data are discoverable by the public, but the data are only accessible/downloadable by data consumers that have been authorized by the data owner. A search would reveal the locations of all variables, sites, and data series, but data consumers would not be able to download the data unless they are authenticated and have been authorized by the data publisher/owner.

Example 1: An academic investigator has received some observations of diversion flows from a private canal company and is storing them on his/her HydroServer. The canal company has asked him/her not to allow unrestricted access to the data because they believe that their data is somewhat sensitive and want to control and keep track of who has accessed the data.

Example 2: An investigator has collected some data and is in the process of quality controlling it on his/her HydroServer. Data QA/QC for the dataset is complete for prior years and those data are ready to be released, but QA/QC for data in the current year is not complete. The investigator wants to control distribution of the raw data to those that are working within his research group. The investigator wants people to know that the data exist, but may not be ready for unrestricted access/download. The investigator also wants to track who has downloaded the data to establish impact as well as for notifying data users when data are finalized.

1. Type 4 – Restricted Data: The metadata are private, the data are not discoverable by the public, and the data are only accessible/downloadable through direct communication with and with permission from the data owner. Data download will only be available to data consumers that are authenticated and authorized. Data consumers will not know that the data exist without learning about it through personal contact with the data provider/owner. The metadata will not be cataloged in a central metadata catalog.

Example 1: An investigator is in the process of collecting data. He/she is streaming the data into his/her HydroServer from sensors in the field. The data is completely raw with no quality control. He/she wants to use the organization of ODM and the HydroServer tools to process the data. He/she wants individuals within his/her own organization to have access to the data, but is not ready for the general public to know about the data yet. He/she doesn’t want to have to maintain separate ODM databases and services for the restricted data.

# 4. Functional Requirements

The following are requirements for the functionality needed to support access control in the CUAHSI HIS:

## 4.1. Metadata and Classes of Data

Data owners/providers will be able to host and potentially publish of all four types of data described above on a HydroServer. Data owners/providers will be encouraged to make the metadata descriptions of all of their data resources public, but it will not be required. Only public metadata stored in an ODM database and published using a WaterOneFlow service will be available for harvest and cataloging in a central metadata catalog where it can be discovered by potential data consumers.

## 4.2. Integrated System

Data access control will be supported as an integrated part of the existing CUAHSI HIS system components. HydroServer will support hosting of both public and restricted data resources in a single ODM database. HydroServer will support delivery of both public and restricted data resources from a single ODM database through a single instance of the WaterOneFlow Web Services. Data owners will not be required to create and maintain separate databases and services for public and restricted data.

A HydroServer will have a single access control system that is independent of ODM (e.g., access control rules will not be stored in an ODM database). If possible, the access control system will be generic so that it can support access control on any digital object type on a HydroServer and not just data series stored in ODM databases (recognizing that HydroServers already host geospatial data through OGC web service and may host additional types of digital objects in the future).

## 4.3. Registration and Authentication of Data Consumers

For the purpose of controlling and tracking data access, registration and authentication of data consumers will be supported by HydroServer. HIS client applications (e.g., HydroDesktop) will also have to support authentication of data consumers. The use of a single sign-in technology that enables data consumers to use a single user account across all HydroServers is preferable. It follows that:

* It is preferable that user credentials be reusable across HydroServers. Because clients like HydroDesktop will request data from multiple HydroServers, it is preferable that one set of user credentials be used to access data from many HydroServers.
* It is preferable that user credentials be globally unique. If user credentials are not globally unique, it would be possible for a person to be unable to get the same user name on every server. For example, a user creates an account with a user name of “jeff” on one HydroServer and then tries to create an account with a user name of “jeff” on another HydroServer only to find that there is already a user named “jeff.”

## 4.4. Authorization of Data Consumers

HydroServers must have functionality that enables data consumers to request authorization to access specific data resources on that server. This may occur automatically through web service calls, or it may involve a web page interface where requests are logged. Data authorization requests will require agreeing to a data access agreement. The data access agreement will specify the terms of data access that is granted. It follows that HydroServers must have functionality that enables data owners to grant or deny access for specific data consumers who have made requests to access specific restricted data resources.

Data consumer access to data resources on any given HydroServer will depend upon which privileges have been assigned to a data consumer. The following rules will be followed:

* Requests from unauthenticated data consumers will be supported by the system, but will be limited to reading public metadata and data.
* Requests from authenticated data consumers will be limited to reading public metadata and data as well as reading restricted resources for which they have been granted authorization.
* In anticipation of create, update, and delete functionality (e.g., via web services or web based data loaders and editors) being developed for HydroServers, such functionality will be limited to authenticated and authorized users (i.e., an anonymous user cannot upload, edit or delete data).

Authorization of a data consumer to access a data resource will be done by the data owner or a HydroServer Administrator. Access control rules for a particular data resource will be stored on the HydroServer on which that data resource is hosted. Software that implements these rules will also be part of the HydroServer on which the data resource is hosted. This requirement implements the principal of local control over the data. If access control was enforced somewhere else the data owner and server administrator would not have the ability to ensure that their access control requirements are being met.

## 4.5. Granularity of Access Control

For time series data, the granularity for access control will be at the series level. A data object to which access can be granted will consist of a data series identified by the unique combination of Site, Variable, Method, Source, and QualityControlLevel (e.g., the results of a WaterOneFlow GetValues web service call). The system will enable setting separate access control restrictions for metadata and data (e.g., public metadata with restricted data). Because of this granularity, the following will be possible:

* Controlling access to both metadata and data separately
* Controlling access to data resources based on their version or quality control level
* Controlling access to all series at a site
* Controlling access to all series for a variable
* Controlling access to all series within an ODM database

## 4.6. Support for Multiple Types of Operations

We anticipate that HydroServer will eventually support create, read, update, and delete operations on data resources through a web service interface. Because of this, the access control system will support authorization of all four types of operations, even though currently only read access is possible. The requirements in this document are only for the access control system that would enable authorization for create, update, and delete operations and do not specify the functionality of the services that would be required to provide create, update, and delete operations.

## 4.7. Use of Web Services

Data access requests and user identity information must be passed through web services. The CUAHSI HIS is a distributed Services Oriented Architecture (SOA) with many different HydroServers. Data is obtained through automated web service calls made on behalf of a data consumer by a client application. New functionality eventually to be supported by HydroServer involves passing create, update, or delete requests through the HydroServer web service interfaces. HydroServer services do not have a visual log-in interface. Because of this, data consumers must be able to identify themselves through web service calls.

## 4.8. Logging Data Access and Download

Logging of data access/download for the purposes of tracking data use and for establishing the broader impact of data resources will be supported. Logging will occur on individual HydroServers and will entail recording the following types of information:

* Who requested the data?
* What type of request was made (e.g., getvalues, getsites, getvariables)?
* Which data resource was accessed?

Logs will be easily accessible to data owners/providers so that they can be queried and summarized. The WaterOneFlow web services currently log data requests and send that information to the San Diego Supercomputer Center. This functionality will be retained so that CUAHSI can generate statistics about overall system use. Individual HydroServer Administrators will be able to disable reporting to SDSC if they so desire.

## 4.9. Data Use/Access Agreements

Data owners/providers will be able to have and use a data use/access agreement(s). The standard CUAHSI Data Use/Access agreement will be provided as a default. Data consumers will have to agree to the terms of this agreement during the registration process and prior to being authorized to access data resources to which the agreement applies. Separate data use agreements may be used for each ODM and WaterOneFlow Service supported on a HydroServer. Data owners/providers should ensure that the data use agreements that they use notify data consumers that their use of the data will be tracked and what particular uses they intend to make of any data tracking information.

## 4.10. Firewall Exceptions

Exceptions in firewalls to allow outgoing communication of data and incoming requests for data should be minimized. To avoid problems with firewalls and to enable HydroServer to be more interoperable with other systems, HydroServer web services (e.g., data access, authentication, authorization) should use standard ports that are not likely to be blocked by University network administrators.

## 4.11. Use of Existing Functionality

If at all possible, the HydroServer access control system should adopt and use existing 3rd party systems (e.g., OpenID, OpenAuth, or other existing systems).

# Acknowledgments

This document was created from input solicited from data managers and scientific investigators working within the Inland Northwest Research Alliance (INRA) Constellation of Experimental Watersheds (ICEWATER). Additional ideas came from Dan Ames, David Valentine, and others working on the CUAHSI HIS project.