Requirements for Supporting Access Control for Published Point Observations Data Using the CUAHSI HIS

Jeffery S. Horsburgh

3/18/2010

1. Introduction

Academic researchers who are collecting data within experimental watersheds, observatories, and research sites have need for the ability to support both private and public data 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 supporting both public and private data, new functionality is needed that enables HydroServer administrators 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 currently no restriction on or logging of who can access the data through the WaterOneFlow web services. There are a number of situations where academic 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. Integrate their data organization, management, and publication rather than maintaining separate systems for each of these functions.
2. Have better control over how, when, and if data go from private to public.
3. Publish research results (e.g., peer reviewed publications) based on data before the data are released to the general public.
4. Keep track of who is downloading and using their data and evaluate its impact on the community.
5. Have and use a data use/access agreement and ensure that they get credit and appropriate citation for the data that they publish.
6. Control who can access/download data.
7. Only expose the best or highest quality data, while restricting access to preliminary or raw versions of the data.

The following sections provide more details about the types of data that need to be supported and the functionality that is needed.

2. Definitions

To facilitate this discussion, some definitions are helpful:

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 use data hosted on a HydroServer.

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

Data Provider – The person or organization that his hosting and managing the HydroServer.

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

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 proposed:

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 registration or 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.

2. 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. Registration and authentication of data consumers would be required for download, but only so data access/use could be tracked. The general data use agreement would indicate the conditions of use for the data.

   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.

3. 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 with permission from the data publisher/owner. A search would reveal the locations of all variables, sites, and data series, but data consumers cannot download the data unless they are registered, 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.

4. 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 publisher. Data consumers will not know that the data exist without 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 Data Types

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 harvested and cataloged 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 private data resources in a single ODM database. HydroServer will support delivery of both public and private data resources from a single ODM database through a single instance of the WaterOneFlow Web Services. Data owners/providers will not be required to create and maintain separate databases and services for public and private data.

4.3. Registration, Authentication, and Authorization of Data Consumers

For the purpose of tracking data access, registration, authentication, and authorization of data consumers will be supported. Data consumer access to data resources on any given HydroServer will depend upon which privileges have been assigned to a data consumer. Authorization of a data consumer to access a data resource will be done by the data owner/provider, and access control rules will be held on each individual HydroServer. The use of a single sign-in technology that enables data consumers to use a single user account across all HydroServers is preferable. HIS client applications will have to support authentication of data consumers.

4.4. Transition of Data from Private to Public

Versioning of data resources will be supported. Functionality will be provided that enables data owners/providers to transition data resources from private to public. This may include an attribute for a data resource that determines when it becomes public in the case of a data embargo period.

4.5. 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 information about which data consumers accessed which data resources and when that occurred. Logs will be easily accessible to data owners/providers so that they can be queried and summarized.

4.6. Data Use/Access Agreements

Data owners/providers will be able to have and use a data use/access agreement(s). 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. Data use agreements would most likely be administered at either the HydroServer or at the individual service level.

Acknowledgements

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 David Tarboton, Kim Schreuders, Dan Ames, David Valentine, and others working on the CUAHSI HIS project.