Earth Sci Inform (2008) 1:119–129
DOI 10.1007/s12145-008-0015-2

RESEARCH ARTICLE

Availability and coverage of hydrologic data in the US geological survey National Water Information System (NWIS) and US Environmental Protection Agency Storage and Retrieval System (STORET)

Bora Beran · Michael Piasecki

Received: 23 April 2008 / Accepted: 23 September 2008 / Published online: 8 October 2008

Abstract The need for a unified and improved data access system for the nation's vast hydrologic data holdings has increased over the past few years as researchers strive for better understanding the human impact on the nation's water cycle. Large mission oriented data repositories such as the USGS' National Water Information System (NWIS) and EPA's Storage and Retrieval System (EPA STORET) play a crucial role in providing a substantial amount of the nationwide coverage, however they do differ regionally in terms of coverage (parameters) and geospatial data density. Besides the differences in geographic distribution, repositories tend to undergo changes in mission statements and as such have different foci in their data collection activities that change as time progresses. This paper places the two water information systems next to each in an attempt to work out the differences in terms of coverage and content and how they complement each other when overlaid. This is done through the use of a number the CUAHSI Hydrologic Information Systems components, namely a web-service suite called WaterOneFlow that permits interrogation of the available data content of a national water metadata catalogue into which these two information systems have been integrated.

Communicated by: H. A. Babaie

B. Beran
Microsoft eScience Group,
835 Market Street,
San Francisco, CA 94105, USA
e-mail: borabe@microsoft.com

M. Piasecki (✉)
Department of Civil, Architectural and Environmental Engineering, Drexel University,
3141 Chestnut Street,
Philadelphia, PA 19104, USA
e-mail: Michael.Piasecki@drexel.edu

Keywords Databases · Data Coverage · NWIS · STORET · WaterOneFlow · Web-services

Introduction

The data spectrum that hydrologists are in interested in by nature is extremely diverse and tends to span multiple time and spatial scales. As a result, it is difficult to actually define a precise limit or threshold that would identify data to be important for hydrology versus labeling it irrelevant. In addition, the ease of access to the data sources as well as the need to having to visit several of them to acquire all the data necessary to conduct an investigation, typically has been (and continuous to be) quite time consuming. Not surprisingly, in 2004 a survey carried out by the Consortium of the Universities for the Advancement of Hydrologic Science Inc., CUAHSI Hydrologic Information System project, HIS (Bandaragoda et al. 2006), returned, among many desires, as the number one wish formulated by hydrologists: "…better and easier access to hydrologic data."

In response to this initial survey and after continued consulting of the hydrologic community (mostly academic at the time, but now more and more embracing other disciplines as well as private and governmental sectors) CUAHSI-HIS has embarked on developing cyberinfrastructure, CI, to respond to the voiced needs (Tarboton et al. 2007, Beran and Piasecki 2008, Zaslavsky et al. 2007). While the envisioned suite of CI developments is quite extensive (Maidment 2005), and concerns modeling, data access, data description, data storage, and higher end concepts like the development of a purely digital representation of a watershed including all its data contents as well as processes, the initial focus has been to work on a "WaterOneFlow" web services environment (Whiteaker et

al. 2007) that permits a programmatic access to a national water metadata catalogue via the internet. Current web-services include a getSites, getParameter, and getValue procedure that permit querying and retrieving of national, regional, and local data sets that can be consumed in end user application using a standard return format (WaterML) encoded in XML.

A central component of the CUAHSI HIS developments is the point observations data model which in turn has been realized as a relational database (Tarboton et al. 2007). This data base model is used for the central water metadata catalogue (compiled and hosted at the San Diego Supercomputer Center) and acts as the underlying data information storage. The definition of the web-services and the standardized return messages (in WaterML format) is beyond the scope of this paper, but suffice it to say that these web-service signatures are designed such that they mimic each other regardless the data source. These web-services in turn can be invoked by back end applications, for example Hydroseek (Beran 2007) which was developed to provide uniform access to multiple local/national hydrologic data repositories.

While hydrologic data collection efforts are carried out by many entities, be it federal agencies, state or local agencies, NGOs, or individual investigators, the federal agencies typically have the largest scope in terms of number of different variables collected, and also their temporal and spatial coverage. Because it would be impossible to examine all data collection efforts, this paper focuses on the two biggest data stewards; the US Geological Survey (USGS) and the Environmental Protection Agency (EPA) which are responsible for the majority of the hydrologic and water quality data collected in the United States. EPA's Storage and Retrieval System (STORET) (www.epa.gov/storet/) and USGS' National Water Information System (NWIS) (http://waterdata.usgs.gov/nwis) make this data accessible through their web interfaces. While USGS is largely responsible for its own data collection efforts (however, it does receive finds from the US Corps of Engineers to collect data on USCOE's behalf), EPA mostly receives data from other entities like state environmental agencies and tribes and thus largely acts as steward and manager for other collection efforts. However, because submission to STORET is not mandatory, not all the local/state agencies participate in STORET due to differences in business requirements. Those areas can be identified as bald spots in these agencies' geographical coverage.

While there are overlaps in data collection activities, these two agencies have different foci, which reflects their different mission targets. STORET is a repository for predominantly water quality data (it does have some flow data also) while NWIS stores water quality data in addition to groundwater and stream data. Hence the two data repositories overlap on the water quality column and complement each other on stream flow and groundwater data. There have been several efforts by the two agencies to integrate their repositories at different levels over the past 6 years. EPA's 'Window to My Environment' (WME) (http://www.epa.gov/enviro/wme/background.html) displays stations in both STORET and NWIS system on an interactive map. However WME does not deal with the heterogeneity problem. Thus for data discovery and retrieval the user is redirected to the respective data provider's website. Following WME, Lockheed Martin Information Technology started a study for EPA and USGS (National Water Quality Monitoring Council Meeting Minutes, Durham, New Hampshire, July 26–28, 2005 http://water.usgs.gov/wicp/acwi/monitoring/minutes/nh_072605.html) which involves associating measurements of the two agencies with one another with the goal of solving the problems related to semantics.

The HIS CI activities first focused on the development of data access systems which necessitated the creation of extensive data source catalogs (a collection of detailed information about what the databases contain). These were compiled from both NWIS and STORET (and other much smaller data collections) using automated programmatic web-page parsing techniques. While the primary goal of harvesting was to create search engine indexes, it also made it possible to gain a deeper insight into these systems by examining the metadata and to deduce some basic understanding about the current content, spatial and temporal coverage, distinct differences between the systems, and the limitations of the two data sources.

The compiled catalogs, besides serving as a key to the WaterOneFlow environment, also provide considerable amount of information on the history of hydrologic data collection activities, data availability and effect of policy making on science. Both agencies deserve much praise for their willingness to cooperate in this endeavor. Several months after the initial harvesting of the NWIS site, USGS agreed to provide CUAHSI HIS developers with the data directly; which, besides making it faster, provided considerable relief from creating incorrect results due to errors created that were due to inconsistent contents in the respective databases. Ongoing discussions with EPA's STORET team (status Spring 2008) will hopefully result in a similarly straight forward mechanism to retrieve and harvest STORET.

Methodology

In order to execute the proposed study three steps were taken. First, the metadata catalogues for both databases

were generated by using database dumps that were received from EPA (STORET) and USGS (NWIS) and then scattered into the underlying observation database model (ODM) metadata database. This step also included a sweep at correcting and supplying incorrect and missing metadata information on stations and their content. Second, the metadata catalogue was also stored in an OLAP cube (see later) for making it possible to rearrange the tables containing the metadata corresponding to desired queries and views. Third, the ODM metadata catalogue was then interrogated using the WaterOneFlow webservices described earlier to extract site data and content and coverage information.

One important aspect when attempting to place databases side-by-side for comparison purposes is the consistency of the information (or metadata) contained in both databases. The consistency of metadata is in fact a necessary prerequisite in order to be able to make comparisons between different data repositories on the basis of data availability, site distribution, data priorities and to aggregate the results to see a better overall scenario. Not surprisingly however, NWIS and STORET are not fully consistent as they both use different terminologies to identify measurements and site types and also exhibit inconsistencies concerning the full availability of geographic identifiers (like a latitude/longitude pair). This in turn prompted the need to investigate and then fill-in some missing information for both (NWIS and STORET) metadata catalogues before the interrogation could commence. For example, Table 1 shows the availability of geographic identifiers for stations in EPA STORET as of December 2006.

From Table 1 it is clear that about 500 stations in the database while containing data cannot be discovered using a latitude/longitude based search box. The majority of these stations have been migrated from Legacy STORET. These stations are often geo-referenced by descriptions like; "CORNER OF CANAL NEAREST INTERSECTION OF 42ND AVE AND 37TH ST S, ST PETERSBURG". Some 800 stations cannot be searched based on state–county information, while approximately only half of them actually have an accompanying Hydrologic Unit Code (HUC) stored alongside the site information. While 500 stations may not seem a lot compared to a total of approximately 274,900 it may be just those one or two stations that would have held crucial information. In addition, state information can be provided using abbreviations or state names or Federal Information Processing Standards (FIPS, www.nist.gov/itl/fipspubs) codes. The NWIS in contrast uses all three state, county and country FIPS codes where applicable. Geographic identifiers as used in Table 1 are available for all NWIS sites except about 500 that are missing geographic coordinates, about 1,800 that do not have state/county information, and 79,192 stations, out of approximately 1.7 million stations total, that are missing hydrologic unit codes (HUC) as of April 2007. Yet, it serves to demonstrate that these undoubtedly very extensive and valuable databases are not free of inconsistencies and missing data pieces. Missing geographic identifiers were computed using ancillary information. HUCs and states were assigned based on coordinates while sites in the ocean or estuaries were associated with the nearest state. When coordinates were not available, site and/or local data collection agency names were used to identify the state the site belongs in.

Table 1 Availability of geographic identifiers for stations in EPA STORET

Total number of sites	274,918
Sites with geographic coordinates	274,435
Sites with state/county information	273,113
Sites with hydrologic unit codes	128,646

In order to have a better and more manageable framework for comparison we have grouped STORET and NWIS into categories based on their primary site types and by combining categories when necessary. Sites were classified into seven categories, namely; stream/river, groundwater, lake/reservoir, estuary, coastal, meteorological and other. Since all repositories have different names for station types and use different levels of detail for classifying them, aggregating sites required reconciliation of these differences. To reduce the number of categories to a level such that they become comparable EPA and NWIS sites were grouped according to their primary site types and by combining categories when necessary. For example EPA's "Well" site type, as well as NWIS' "Groundwater" and "Spring" were listed under the category "Groundwater". "Other" represents sites in wetlands, facilities (e.g. treatment plant effluent) and man-made drainage and transportation channels. Table 2 provides a list of some mapping examples. The full list is not provided here for brevity's sake since EPA and NWIS have a combined total of 131 site types.

Table 2 Site type mappings used in the analysis

STORET site type	Mapping
River/stream	Stream/river
Great Lake, lake, reservoir	Lake/reservoir
Well, other-groundwater	Groundwater
Canal-transport, Waste pit, wetland	Other
NWIS site type	Mapping
Stream, stream-diversion, stream-outfall	Stream/river
Lake/reservoir+meteorological	Lake/reservoir
Spring, ground-water other than spring	Groundwater
Meteorological	Meteorological
Land application, outfall	Other

A similar heterogeneity problem exists concerning measured parameter names. Both NWIS and STORET individually use approximately 10,000 variable codes to identify their measurements. This means that one can find about 10,000 variables that are common to both systems albeit different using code structures. Approximately 725 parameters in NWIS have no match in STORET, and approximately 850 parameters in NWIS have no match (personal communication with EPA on the Water Quality Exchange effort, WQX 2008).

Because there are thousands of different parameters between the two databases it was decided to use a broader classification system so that a state-by-state comparison including parameter coverage could be carried out with numbers that can be compiled into concise graphics without overburdening the image. The parameters were classified into 10 categories: Streamflow, Stage/Gage Height, Reservoir Storage, Groundwater Flow, Groundwater Level, Water Quality, Soil Quality, Meteorology, Oceanography and Other. Streamflow, Stage/Gage Height, Reservoir Storage, Groundwater flow and Groundwater Level categories contain parameters related to quantity of surface- and groundwater. Water Quality and Soil Quality parameters contain physical, chemical and biological environmental quality parameters and the Meteorology category contains measurements related to precipitation, evaporation, air temperature, barometric pressure, winds, solar radiation and heat fluxes. "Other" represents ancillary data, i.e. parameters such as 'Project Code', 'Sampler Type' and 'Location in cross-section'. In order to be able to analyze the considerable amount of data that is available, Online Analytical Processing (OLAP) (Chaudhuri and Dayal, 1997) technology was employed. OLAP uses a multidimensional data model, allowing for complex queries with significantly reduced execution times. This model supports pre-computed aggregations of records such that summaries based on a certain data attributes can easily be created. Maps were generated using ArcMap based on data from OLAP data cubes.

Measurement sites

The NWIS and STORET systems have a total of approximately 1.7 million stations distributed over the entire nation, also noticing that NWIS has approximately 6 times as many sites as STORET. The site count per state, however, varies greatly from one state to another. For example, STORET has only 863 sites in the state of Texas while 47,602 sites in Florida. On the other hand, NWIS has 27,906 sites in Florida but 121,545 in Minnesota which is almost 5 times the number that STORET has in Minnesota. Figure 1 shows a color coded distribution of sites per state for STORET, NWIS and the union of the two systems.

In some states NWIS and STORET seem to complement each other well in terms of site numbers, as in Arizona and even more so in Minnesota and Florida, while other states in general seem to have a weak coverage despite the union like New Mexico, Vermont and a good number of Mid-western states. This however should not be misinterpreted as a lack of interest by either system to host monitoring data nor should it be misinterpreted as a lack of interest in monitoring environmental parameters by these states. For example, the reason for EPA's weak coverage of Texas is that the only data contributors to STORET are the National Park Service, EPA National Aquatic Survey, Texas General Land Office and The Rivers of Colorado Water Watch Network. In contrast, the Texas Commission of Environmental Quality (TCEQ), a major collector of environmentally relevant data in the state of Texas, does not participate in STORET but disseminates its data using a local data distribution system. In fact TCEQ has 18 times more water quality data than EPA STORET in Texas. In essence, the coverage shown in Fig. 1 is not a reflection of state bias by either STORET or NWIS, but merely a reflection of what volume of data a researcher can expect to find in any given state by just using these two databases, i.e. Fig. 1 favors larger states since it shows the site count rather than the density. Figure 2 provides a different view in that it presents site densities measured in sites per square mile per each state to provide an alternative way and more informative representation. While this approach is perhaps more intuitive it should be kept in mind that density of the coverage is also a function of terrain and stream network.

The numbers presented in Fig. 1 are only for the 48 contiguous states of the United States and considerable coverage exists for Alaska, Hawaii, and Puerto Rico, as shown in Fig. 3. Also, coverage is not limited to the US but data exist for countries such as Canada, Mexico, Ukraine, Japan, Afghanistan, Iraq and several islands in Pacific and Caribbean, even though the number of sites for each of these states is considerably lower.

Using the classification scheme (for sites) outlined earlier all sites can be placed in the 7 classification bins to provide a better understanding how many sites are actually dedicated to collecting data within each of these groups. Figure 4 shows different site types for STORET and NWIS and it is evident that the number of sites within each bin group varies dramatically from each other. NWIS collects groundwater related data from about 90% of its sites that together with STORET, comprise about 75% of all 1.7 million sites. In contrast only about 10% of all sites fall into the Stream/River category, and only 3% to Lakes/Reservoirs. This seems to suggest that there should be an abundance of groundwater data while stream/river related data is already sparse when compared to groundwater.

This also seems to suggest that from a research and science interest point of view groundwater figures promi-

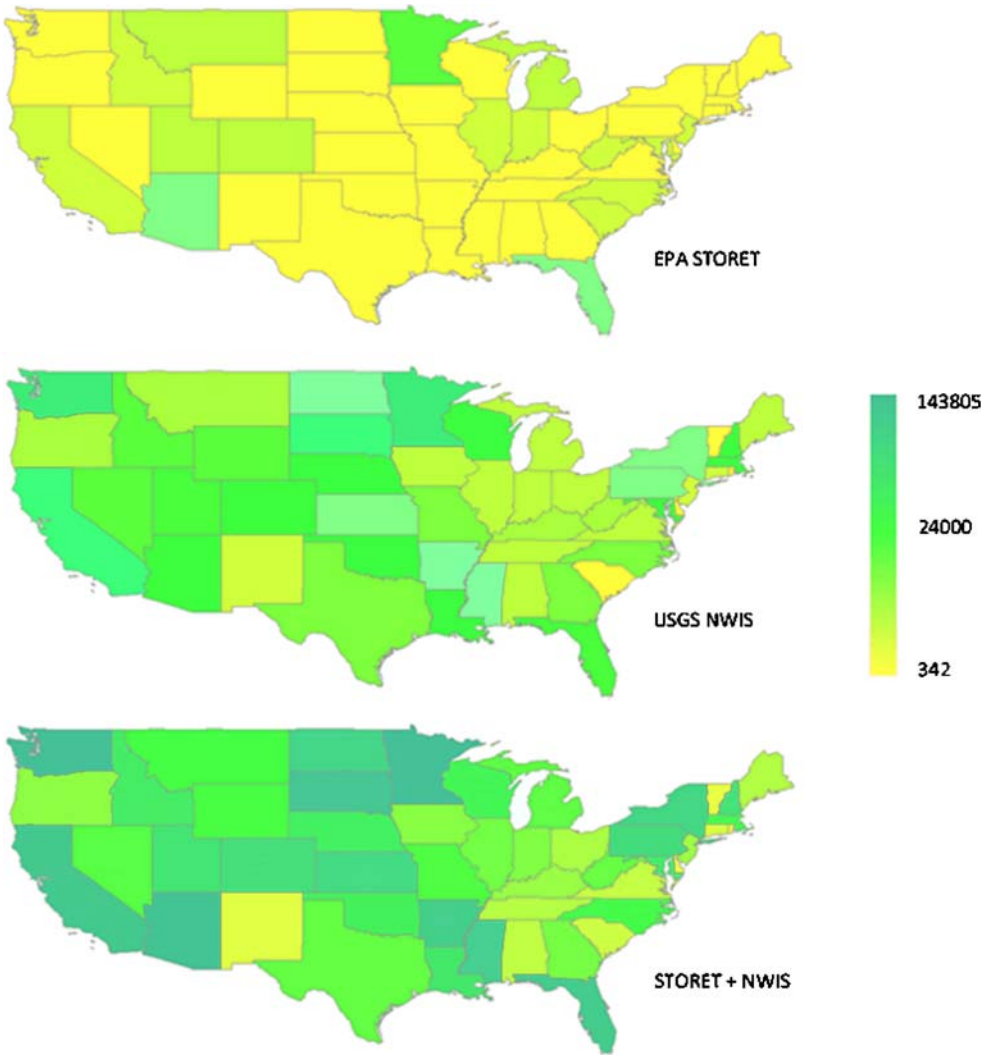

Fig. 1 Geographical distribution of STORET and NWIS sites across US

nently on the science agenda (particularly within USGS), while all other water bodies suffer from neglect or disinterest. This is of course not true as is it also important to take a close look at how many data channels each site operates, i.e. how many different parameters are being collected at the same site and thus what the data diversity is. Last but not least it should not be forgotten that different environmental monitoring needs require different densities of instrument distribution as well as sampling or measurement intervals. Environmental processes (hydrologic ones being a subarea) take place at different temporal and spatial scales that vastly influence the number of stations necessary to achieve a certain monitoring objective. In view of this it is perhaps clear why NWIS features so many groundwater

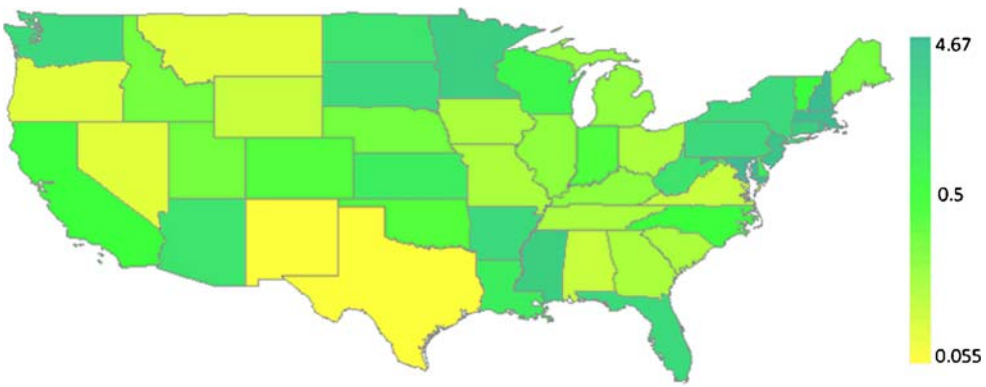

Fig. 2 Sites per square mile for NWIS and EPA STORET

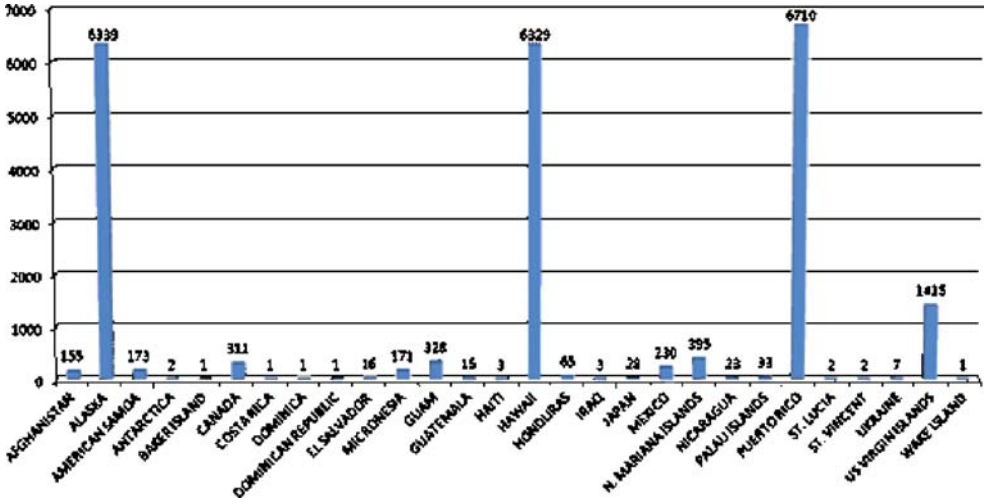

Fig. 3 Total number of sites not visible on the map of continental US

well sites, because hydrologic sub-surface processes are difficult to access, they happen over large spatial and temporal scales, and they can exhibit substantial variation over smaller scales all of which contribute to the need to having a relatively dense array of wells. The question remains however, how much does the groundwater bin contribute to the overall data amount.

Data availability

Figure 4 suggests that groundwater data makes up the bulk of the total data available. However, a station can have several collecting channels, i.e. monitor or collect samples for a variety of parameters. If one is interested in the volume of available data and also its variety a better indicator is to simply look at the number of available records per site. A record in this context means a data value collected, this also takes into account the fact that some measurements may be carried out every 10 or 15 min while others are collected every hour or even be conglomerated into a daily value only. Figure 5 shows data availability for each network based on data values on record.

The combined holdings of NWIS and STORET comprise about a total of 350 million data points most of which is stream flow data from NWIS. It also shows that groundwater data is only a small portion of available NWIS data and that EPA and NWIS have a similar amount of water quality data. However, it is clear from the right panel that the mission of STORET is clearly focused on collecting water quality data (it comprises about 95% of all data holdings inside STORET), while NWIS has a somewhat broader data mission even though about 75% of the data is related stream flow. Figure 6 provides the distribution of the 350 million data points between the states. Since data point count is also a function of length of record (period of time over which data was or is collected at a site) and not only the site count, this figure differs significantly from Figs. 1 and 2.

Above graphs document what both NWIS and STORET actually hold in terms of pure storage. However, even though all sites are accessible and will appear on the data access

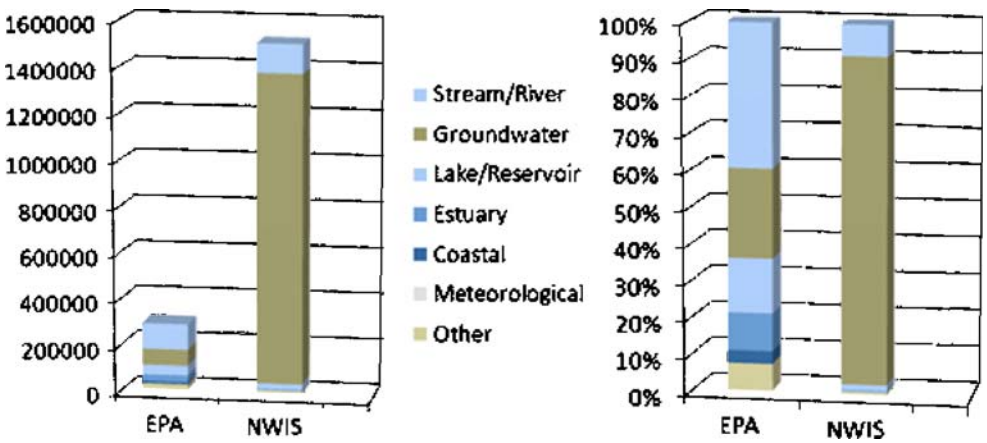

Fig. 4 Site type distribution for EPA STORET and NWIS

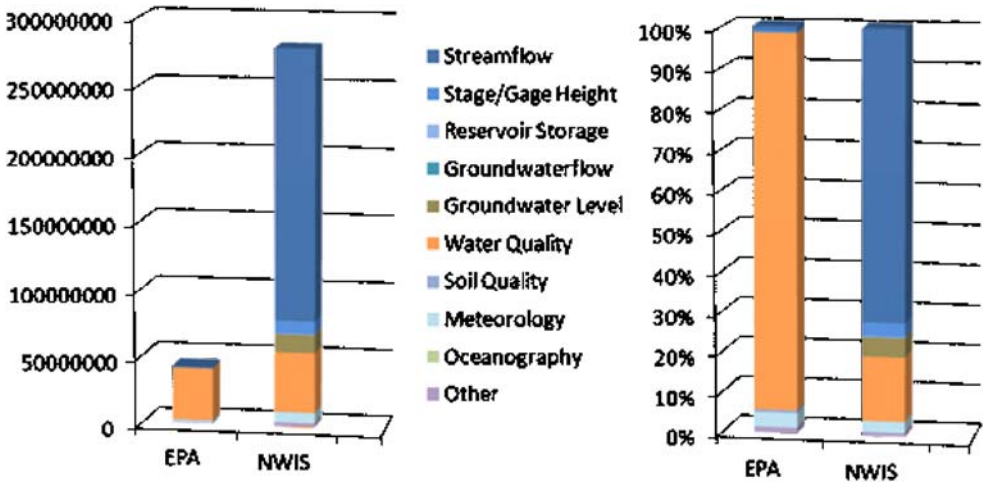

Fig. 5 Data availability for EPA STORET and NWIS

portals when queried, many do not offer up any data. One such site is the groundwater monitoring site 374028101001001 (located in Haskell County, Kansas, http://waterdata.usgs.gov/nwis/inventory/?site_no=374028101001001), which simply states that no data is available and that the state office needs to be contacted. Hence, while data is present it is not necessarily available to the public without special request. This distorts the data assessment and availability considerably (as presented in Fig. 4) and also somewhat explains the disproportional number of groundwater sites when compared to the total data they hold. Hence, a new analysis was conducted to extract only those sites that actually have something to offer when queried.

It is of interest also to examine the question of how many sites are actually providing data, i.e. having at least one data point recorded. To answer this question we examined Period Of Record (or POR) tables that were received from the NWIS data management team as well as the STORET data base dump. Table 3 shows that of the approximately 1.5 million sites available in NWIS about 1 million have a data record. Of these, about 370,000 have water quality data, 813,000 contain groundwater data, and 62,000 have physical characteristics data (mostly stream-flow and gauge data). While groundwater sites are still a considerable portion of the total number, the NWIS site count is about 500,000 less than it was in Fig. 4. Moreover, a similar change can be seen for EPA STORET site count as well. For example, the Florida Department of Environmental Protection has 2,926 sites in the STORET system, 2,379 of which offer no data. In total about only half of the STORET sites (approximately 135,000) actually offer at least one data value. As a result, sites that are usable add up to approximately 1.17million (combined NWIS and STORET), which are significantly less than the total number of sites reported earlier, i.e. about 1.7 million. Notice that both NWIS and STORET hold about the same number of water quality records,.i.e. about 4.6 million each, despite NWIS having more than twice the number of stations recording water quality. Finally, the real strength of NWIS lies in its recording of surface and groundwater data. For example, the 813,000 GW sites in NWIS have about 8.2 million data records (a record being a recorded data value), an average data density (#of_records/#of_sites) of 10 records/site. Note that this is an average value (and s such a guide only) computed from all data records available (earliest: January 1900; latest: May 2008). The majority of

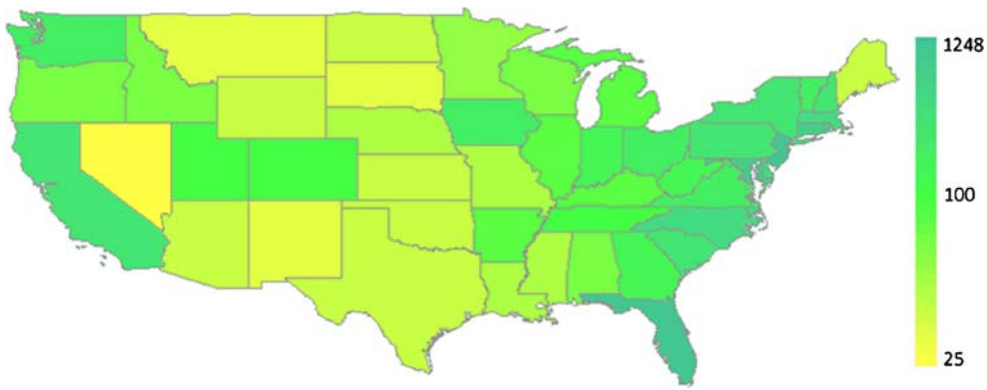

Fig. 6 Data points per square mile for NWIS and EPA STORET

Table 3 Available data sites and average data densities in NWIS and STORET

	NWIS			STORET		
	Number of sites	Number of records	Av. data density	Number of sites	Number of records	Av. data density
Water quality	~370,000	~4,660,000	12.6	136,000	4,500,000	33
Surface water	~63,000	~30,000,000	476.2	small	small	
Ground water	~813,000	~8,200,000	10.1	N/A	N/A	

sites (about 90%) has only a small number of records (<10, actually 75% have only a single value) and only 14,200 sites (<2%) having more than 100 records. If one counts only those sites that have more than 10 records the average number of values recorded per year for those sites is about 14.2. Here too, many sites have only about 1 value per year, with some only 1 value per 3 years, while others have 1 value per month, and some even 1 value per week for the period of records covered. Finally, the 63,000 surface sites hold approximately 30 million records, an average data density of 477 records/site.

History

As the expansion, maintenance, and operation of sites is a cost intensive effort that is subject to budgetary constraints encountered by both institutions, it is helpful to take a look at how the number of sites per state have evolved over the past two centuries, i.e. from the time when the first data was recorded. This is not to say when EPA and USGS started recording data or when STORET or NWIS were set up (obviously these institutions and databases were established much later) but going back to data records as old as the 1800s. The number of sites and the trend is also an indicator of increased awareness of the utility of monitoring data particularly in view of the degree of industrialization over the past two centuries and the resulting environmental impact this has caused. Figure 7 shows a map (NWIS and STORET combined) of when data started to be collected and its progression through time.

It is clear from Fig. 7 that collection of environmental data was literally nonexistent from 1800 to 1875, Colorado being the only state having data that reaches back that far. While coverage increased to encompass the entire continental US by 1900. Even by 1925 density, i.e. sites per state, was still very low—a clear indicator that during the period of increased industrialization, development, expansion, and settlement of the continent, little effort was spent on data collection programs that would monitor environmental health. Even by 1950 the density of sites has not increased by much, while the period from 1975 to 2007 has seen a dramatic increase in operational sites, no doubt also a result of legislative work like the Clean Air Act (1963), and the Clean Water Act (1977). Notice that the above figure, while certainly indicative, is not entirely accurate for reasons mentioned before: states like New Mexico, Oklahoma, Vermont and Maine (or even Texas) seem to have received less attention which is not necessarily true as they may have organizations that, while having substantial data collections, have not contributed to either NWIS or STORET (like the TCEQ in Texas).

In Fig. 7 numbers have been compiled based on cumulative records. In other words, sites that are no longer operational (for example, the most recent estimate of the number of long-term USGS streamgages discontinued since 1990 is now 725) still count since they offer historical data, which may be of value. As a result, these plots do not show

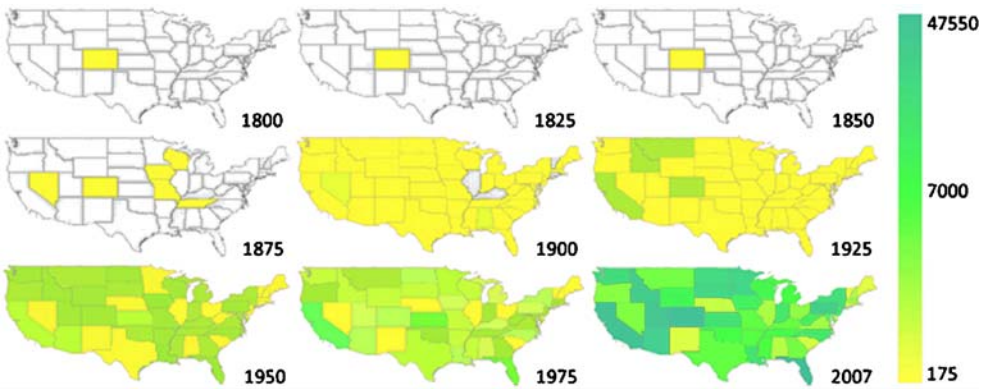

Fig. 7 STORET and NWIS monitoring sites over time

the number of active sites in a given year which is a better measure for data collection activities in any given year. Figures 8 and 9 show the number of stations over time for EPA STORET and NWIS respectively in which a site is considered active from the earliest measurement date until the final measurement date. Note that sites with life spans under 1 year are not displayed in this figure.

It is clear from both figures that the number of active sites has decreased over the past 20 years (for NWIS) and 10 years for (STORET). Even though some of the funding for maintaining gages comes from cost-sharing partnerships between the USGS and more than 800 state, local, tribal, and other agencies, (Christen 2005) and EPA STORET having 305 similar partnerships, the primary reason appears to be that budget allocations have not kept up with needs to maintain all sites. While some fluctuations in the number of active sites are not unexpected, both USGS and EPA funds allocated for operating and maintaining monitoring sites have decreased (Ursery 2004; Renner 2007; Stokstad 2001; Bush 1992; Showstack 2004, Rogers, 2001) despite small and continuous increases in the overall budgets. For example, the water portion of the USGS budget has seen a slight increase of 0.6% (from 202.83 million in FY 2002, to 207.15 million in FY 2003), (from USGS Office of Budget http://www.usgs.gov/budget/funding_tables.asp). However, mandated pay raises in excess of 0.6% have effectively reduced the available funds for the monitoring programs.

While other reasons may have contributed to the decrease in available funding like the fact that some sites may have fulfilled their "duty" in collecting data, or a realignment of monitoring sites, or the addition of strategically placed sites that could have resulted in eliminating a larger number of sites previously needed to achieve the same monitoring objective, the primary reason is clearly budget related. In view of increasing demands for environmental protection (for example the Total Maximum Daily Load, TMDL, development), the need to look into the better understanding of interrelated processes at larger time and spatial scales, impacts of global warming and associated environmental changes on all scales, seem to suggest that this trend points into the opposite direction from where it should be heading. While an effort like that of CUAHSI to bring together as many data sources into a single data access environment and thus providing a means to amend the data contained in STORET and NWIS, it is clear that both NWIS and STORET are two of the most important databases this nation has for surface (and sub-surface) water point-monitoring data. A strong support for these two data stores will remain essential in providing nationwide coverage that serves as the foundation of all other data that can be added regionally to this base set.

Summary

The two largest water data collection networks in the United States (USGS NWIS and EPA STORET) have been examined to better understand the state of data availability, variety and geographical distribution. This work was carried in conjunction with the development of a nationwide hydrologic data information system that seeks to place all (or at least nearly all) available data within one analysis environment. More specifically, it is the result of creating catalogues of data providers that can be interrogated to discover and retrieve data of interest. Because the data horizon is vast, the data was categorized into 10 bins to make analysis and comparison easier. We have found that:

- USGS NWIS has about 6 times as many stations as compared to STORET (1.7 million vs. 274,000). However, most of the stations are dedicated to

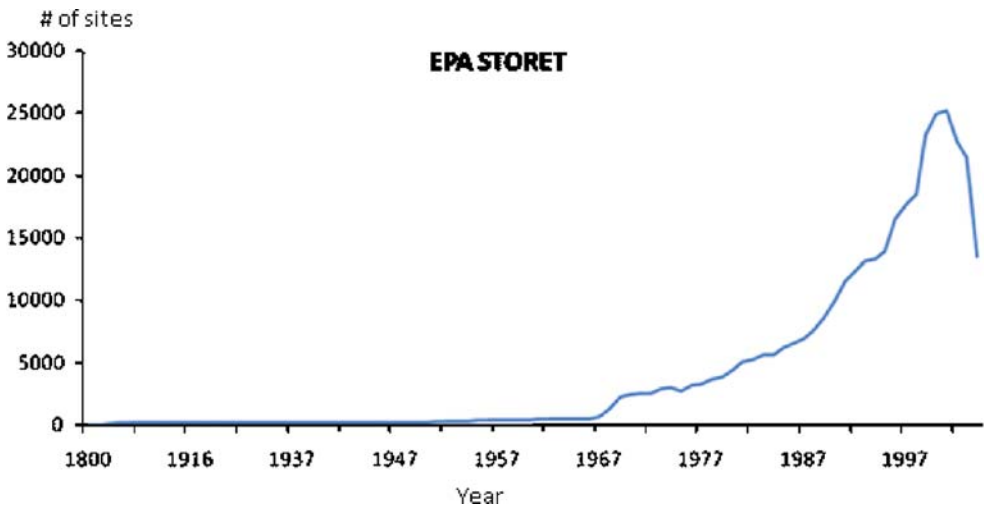

Fig. 8 Active STORET sites over time

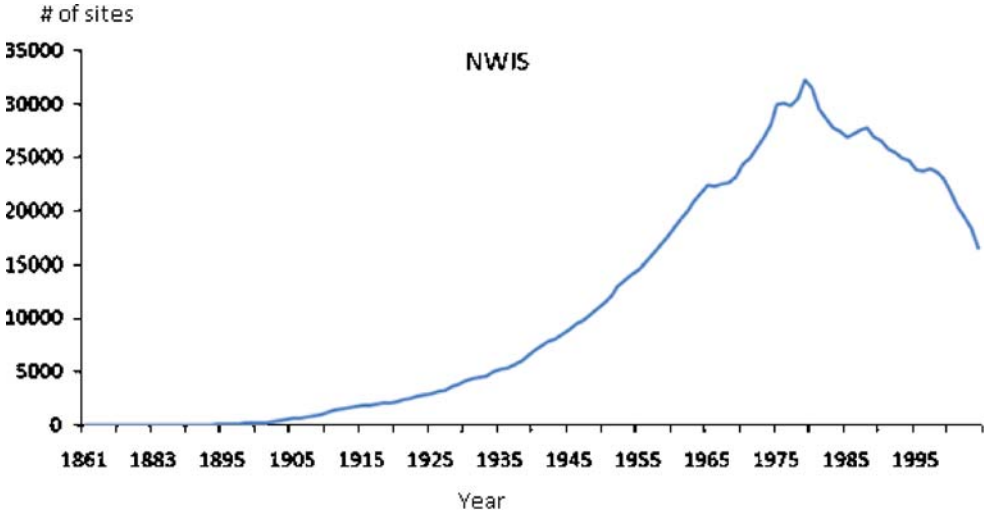

Fig. 9 Active NWIS sites over time

groundwater monitoring (about 90% of all sites) while the remainder is largely dedicated to stream flow monitoring. STORET on the other side has a clear focus on collecting stream water quality data (it makes up ~40% of all sites even more when counting the lakes and reservoirs).

- There is a clearly defined geographical disparity in coverage, i.e. not all states enjoy the same number of sites. While this is not entirely surprising as small states should not necessarily have the same number of sites when compared to states that are 15–20 times larger, it does not fully explain the differences, i.e. similarly sized states still show large differences. USGS NWIS tends to have more coverage in the western states than in the eastern states, while EPA STORET has three states (Nevada, Florida, and Minnesota) that have significantly larger number of sites than the others. However, one needs to realize that low coverage is not synonymous with lack of data, as data may reside in other data centers, like TCEQ in Texas, an institution that does not participate in disseminating their data collections within NWIS or STORET.
- Even though the bulk of sites are related to groundwater monitoring, the bulk of actual data belongs in the stream monitoring bin. In other words, the stream monitoring sites collect a lot more data in terms of variety and also in terms of collecting frequency. 75% of NWIS and STORET combined data is dedicated to stream flow measurements (NWIS holding the bulk of this information). About 20% of the total is related water quality parameters (STORET holding about as many records as NWIS) while the rest is split among the other 8 bins.
- The database structure of both NWIS and STORET is not visible to the public, i.e. the total amount of data and ancillary information is not disclosed. NWIS and STORET each have a clearly defined access portal or window that is open to the public which limits access to a subset of the entire holdings. In fact, 66% of all sites offer up data that can be retrieved. The other 33%, while visible as existing stations, do not allow data querying and data must be requested (however, these are mostly groundwater monitoring sites and only marginally affect the number of stream and WQ sites). This limits the total number of immediately "useful" sites to about 1.17 million sites.
- While records reach back to the 1800s' significant expansion of both databases took place in the 1970s continuing though the 1980s and 1990s. Coverage over time has reached all states in the US, even though regional differences are noticeable, i.e. certain states and watersheds have a fairly dense network of stations while others do not. The peak of available stations was reached in the mid 1990s with approximately 46,000 that were active. However, largely due to budgetary constraints in both host institutions, the number of active (data collecting) sites has continuously decreased over the last 10 to 15 years, a trend that is unfortunate. Particularly in view of increased public environmental awareness and realization that fresh water is not a commodity with endless supplies, one would expect monitoring efforts to increase not decrease. Both databases are viewed in the research communities as the two key data sources for water quality and stream data with nationwide coverage and continuation of these sources is vital for science as well as regulatory purposes. The CUAHSI HIS effort seeks to place these two data bases and their content in the spatial and temporal domain next to smaller data holdings thus hoping to produce a more complete and

comprehensive data space that can serve a base for future expansion with specific foci.

Acknowledgements The work presented in this paper was sponsored by the National Science Foundation through grant numbers 0609832 and 0412904. The authors would also like to acknowledge the Consortium for the Advancement of the Hydrologic Sciences, Inc. (CUAHSI) for their support and contribution in developing the ideas in this manuscript.

References

Bandaragoda C, Tarboton D, Maidment D (2006) Hydrology's efforts toward the cyberfrontier. Eos Trans AGU 87(1):2

Beran B (2007) HYDROSEEK: an ontology-aided data discovery system for hydrologic Sciences, Ph.D. Thesis. 155 pp., Drexel University, Philadelphia, 5 September

Beran B, Piasecki M (2008) Engineering new paths to hydrologic data. submitted to *Computers and GeoSciences*, Elsevier, Accepted for publication December 2007

Bush S (1992) USGS budget cuts threaten programs. Eos Trans AGU 73(15):162

Chaudhuri S, Dayal U (1997) An overview of data warehousing and OLAP technology. SIGMOD Rec 26(1):65–74

Christen K (2005) Funding woes eroding national stream gage network. Environmental Science & Technology, http://pubs.acs.org/subscribe/journals/esthag-w/2005/jan/science/kc_stream.html retrieved July 11, 2007

Maidment D (2005) Hydrologic information system status report. September 2005, http://www.cuahsi.org/his/docs/HISStatusSept15.pdf

Renner R (2007) Budget cuts increasingly damaging to EPA. Environmental Science & Technology, http://pubs.acs.org/subscribe/journals/esthag-w/2007/may/policy/rr_EPA.html, retrieved July 12 2007

Rogers D (2001) Bush, seeking to make room for tax cuts, Tightens Budgets for Science Agencies. The Wall Street Journal, February 16, 2001, page A12

Showstack R (2004) USGS budget would decrease and face inflationary pressures. Eos Trans AGU 85(9):89

Stokstad E (2001) USGS braces for severe budget cuts. Science 292 (5519):1040

Tarboton D, Horsburgh J, Maidment D (2007) CUAHSI Community Observations Data Model (ODM), Version 1.0, Design Specifications. May 2007.http://www.cuahsi.org/his/docs/ODM1.pdf

Ursery S (2004) Congressional subcommittee approves EPA budget cuts. Waste Age, http://wasteage.com/news/Congressional-Subcommittee-Approves-EPA-Budget-Cuts-0722041/, retrieved July 18 2007

Whiteaker T, Tarboton D, Goodall J, Valentine D, To E, Beran B, Min T (2007) HIS Document 5: CUAHSI WaterOneFlow Workbook. Version 1.0, http://www.cuahsi.org/his/manuals/HISDoc5_UseWebServices.pdf

WQX (2008) "Water Quality Exchange Standard" defined by the National Environmental Information Exchange Network, accessed August 2008. http://www.epa.gov/storet/wqx.html

Zaslavsky I, Valentine D, Whiteaker T (2007) CUAHSI WaterML. submitted to Open Geospatial Consortium as discussion document, May 2007, http://www.cuahsi.org/his/docs/WaterML-030-forOGC.pdf