The Case for a Corps Hydromet Data Exchange Standard

Stu Townsley
CESPK-ED-D
916.557.7109
mailto:edwin.s.townsley@usace.army.mil

Abstract

With the explosion of Web-based data exchange, nearly all facets of electronic information exchange have been modified to include an "extensible" format – a format that is extendable to meet individual needs. A standard, XML, has already been developed, adopted, and integrated in many of the latest browsers and other software components. The merits of such a standard are clear: (1) It can reduce and often eliminate specific computer platform dependency and specialized data formatting rules. (2) As the Corps increasingly is asked to collaborate with stakeholders, NGO's, and other federal and state agencies, it can reduce the burden of configuring specialized data formats, which consume precious time and limited resources. (3) It can also reduce security concerns: data exchanges can be removed from core computational and database platforms, thereby reducing the need for generalized access to sensitive resources. (4) And it can give the end-user the ability to fully customize data display, which meets the oft-stated criteria for centralized but adaptable security and standards, and for local control and customization.

Specific standards are being developed, and in many cases already exist, for financial, legal, and scientific data. NOAA, for example, has already begun implementing such an extensible format for spatial data, and the USGS is developing HydroML for hydromet data and for information about sensor site visits. The standard exists, it is being adopted by other organizations. It's time for the Corps to embrace it and adapt it to our specific needs.

Introduction

Several years ago, General Motors realized that the average age of its Oldsmobile customer was over 60. Faced with the prospect of its clientele becoming literally extinct, GM adopted a new advertising slogan – "Its not your father's Oldsmobile!" While I am not suggesting that we are faced with such a dire fate as Oldsmobile, all demographic information shows that not only the Corps, but other federal agencies are going to suffer a major brain-drain in the next five years. Half of us are going to retire. Those hired as replacements come from a dramatically different world.

25 years ago, engineers and scientists were just learning how to harness the power of a computer, linear programming languages such Fortran and Basic were just being touted as advancement from machine language coding. That user would enter data on 80-column punch cards and wait hours or days to see output generated from room-sized mainframes. Quasi-quantitative studies – based on SCS curve number for instance - dominated due to the lack of data and/or the inability to perform multiple complex simulations. As the science evolved to take advantage of computing

horsepower, UNIX machines running HSPF became accessible to more of us. Next Apple, then Microsoft and Intel.

Today, new engineers ask why an HEC-1 line of code is called a card; they likely have never worked on a UNIX machine or even had much to do with command line interfaces. I believe we have an obligation, and a responsibility to ensure that the next generation of Corps engineers and scientists have not only the training, but the modern tools necessary to meet our mission – a mission that to an increasing extent relies and demands on partnerships and stakeholders, a mission that is based on the concept that knowledge is power, that open and unfettered exchange of ideas and information will benefit all.

We, representing the Corps H&H teams, have started to embrace a next generation of technology. HEC-1 has evolved to HMS, HEC- 5 to ResSim, WMS utilizes GIS. However, many of us in the hydro-meteorological (hydromet) community still rely on the Standard Hydrological Exchange Format (SHEF) to share data. Other agencies, and industries have already adopted, or are in the process of adopting, a new standard to define electronic information exchange. This standard is XML, eXtensible Markup Language. The on-line Miriam-Webster dictionary defines extensible as “capable of being extended.”[1] In other words, the standard is fluid and adaptable to new requirements and specifications. On-line retailers use it to process orders and inventory, NOAA is using this standard as the basis for coastal geomorphology, DECODES handles meta-data with it. Others in our field are already seeking to define the rules and structure of an XML subset used for hydromet information.

An Established Bureaucratic Framework

On December 17, 2002, President George Bush signed into law H.R. 2458, the "E-Government Act of 2002. This legislation expands the E-Government initiative to in expanding the use of the Internet and computer resources in order to deliver Government services, consistent with the reform principles … outlined for a citizen-centered, results-oriented, and market-based Government.”[2] The Chief Information Officers (CIO) Council which serves as the principal interagency forum for improving practices in the design, modernization, use, sharing, and performance of Federal Government agency information resources, has specifically designated an XML committee that has been “charged with pursuing:

1. XML Best Practices and Recommended Standards
2. Partnerships with Key Industry and Public Groups Developing XML Standards and Specifications
3. Partnerships with Governmental Communities of Interest to Accelerate the Delivery of XML Benefits
4. Results-Oriented Education and Outreach
5. Projects and Products Benefiting Stakeholders of Multiple Agencies”[3]

Gerry Laniak, of the EPA, presented a paper at the Federal Interagency Hydrologic Modeling Conference detailing a Memorandum of Understanding (MOU) among seven Federal Agencies to pursue collaborative research in technical areas related to environmental modeling. He states that:

"Among the primary objectives of the MOU are 1) to provide a mechanism for the cooperating Federal Agencies to pursue a common technology in multimedia environmental modeling with a shared scientific basis, 2) to reduce redundancies and improve the common technology through exchange and comparisons of multimedia environmental models, software and related databases, 3) to exchange information related to multimedia environmental modeling tools and supporting scientific information for environmental risk assessments, protocols for establishing linkages between disparate databases and models, and development and use of a common model-data framework, and 4) to facilitate the establishment of working partnerships among the cooperating Federal Agencies' technical staff in order to enhance productivity and mutual benefit through collaboration on mutually-defined research studies. In direct support of the goals of the MOU this workgroup has been formed with a specific focus of the computer software infrastructure necessary to support state-of-the-science environmental systems analyses."[4]

Standards defining communications, software, and hardware are ubiquitous. Most have undergone revision as supporting technologies have become more powerful, and more accessible. From the President down, the Federal government managers are recognizing that adopting new computer-based technologies can provide better services to the taxpayer, reduce operating costs, and provide for enhanced mission security. Clearly, the foundation has been established for defining a hydromet subset of XML designed to meet the need of the next generation of stakeholders, NGO's, A&E firms and government agencies to easily and quickly share data.

Current Data Exchange Standard – SHEF

SHEF is a documented set of rules for designed for "coding of data in a form for both visual and computer recognition. It is designed specifically for real-time use and is not designed for historical or archival data transfer."[5] Although SHEF was originally intended designed to facilitate electronic data exchange between organizations interested in hydromet data, one of the offshoots was the development of data collection platforms (DCP) and base stations that used the SHEF standard for internal data collection. SHEF was designed in the in the early 80's when the size of data was a controlling and limiting factor. Not only have the sizes of storage and RAM increased, processor speeds have doubled and cost have been halved every 18 months following Moore's Law. In addition, electronic communication continues to evolve at a similar pace. Twisted pairs are T-3, analog phones have evolved to Digital Cellular Data, and satellites are extensively used with increasing data rates. Therefore, one of the primary motives – an extremely compact format - behind the multi-agency agreement on SHEF no longer is an issue.

Figure 1 shows an example of SHEF data from the Sacramento District. Notwithstanding the "visual recognition" design, this format not easily readable without the cipher nearby. A missing character or an inappropriate symbol renders that data unusable.

```
.A SCCW 20030407 Z DH1315/USC 1.6
.A SCCW 20030407 Z DH1315/UDC 149
.A SCCW 20030407 Z DH1315/TPC 74.7
.A SCCW 20030407 Z DH1330/PCC 12.97
.A SCCW 20030407 Z DH1330/EDC 2.09
.A SCCQ 20030407 Z DH1315/HGC 2.54
.A SCCQ 20030407 Z DH1315/TWC 53.3
```

Figure 1. SHEF A-type

Current SHEF standards use a character pair, the PE code, to identify data type. This means that there are only 676 possible data types. While this is generally enough to handle single data collection operations, i.e., flow and water quality for White River basin in Indiana, this structure must accommodate all users. In our situation in Sacramento, we are interested in sharing evaporation data. SHEF codes only define evaporation in general terms while we want to send the lake evaporation data as both a depth and as a volume. The work-around is to define a new PE code outside of the Another problem encountered using SHEF is that many data collection locations have the essentially the same name. For instance, there are several Dry Creeks in California, and probably at least one in every state. Although the SHEF standard allows for 8 characters ID's, one can quickly run out of location mnemonics that have a readily apparent meaning. In other instances, a new sensor is added to a location measuring a parameter already being measured – this can be done for redundancy, or because the location has undergone physical changes necessitating additional range of measurements, e.g., when a streambed degrades below the lowest intake for a stilling well. Cost can frequently prohibit a retrofit, whereas an additional sensor can be added relatively cheaply.

SHEF is not designed to be a database. Excel, Access, Oracle, Paradox, DSS all offer efficiencies in speed and space. Therefore, data must be converted to a variety of formats before use in modeling efforts. Furthermore, as hydrologic and hydraulic software matures, and as the complexity of required studies increases, it is necessary to use the output from one model as input to another. Although HEC had developed the CWMS suite of software based on DSS, The Waterways Experiment Station has developed similar tools that in many cases mimic HEC's efforts. While WMS can use an HEC-1 rainfall-runoff model as part of a modeling effort, other components require either WMS based files or ASCII imports to the appropriate card(s). In addition, our partners in many studies and in real-time operations often choose well-known, but different models, such as Riverware, or EPA's SWMM. And in some instances, specific projects have adopted custom software – the Reservoir Release Forecast Model for Folsom (RRFM) is an example of a joint effort in Sacramento that involves the Bureau and a local agency, SAFCA, and the Corps. The multiple and often incongruent formats necessitate hours of mundane efforts to convert data from one to another format. One current solution to these problems is an extensive set of pre- and post-processors designed to manipulate data. However, as the underlying software is updated, often times the data format requirements can also change. Another is to use commercial off-the-shelf software such as Access as a database because of its familiarity to a new generation of users. In addition, the reformatting task is not generally given to a programmer, but rather an engineer or hydrologist who writes new code to automate data conversion as part of the effort. Not only is this an inefficient use of talent and skills, these translation programs are frequently ad-hoc, developed solely for that specific task. This leads to software that is not portable to another study, and lacks the requisite level of documentation for archival purposes.

Why XML?

Most organizations have discovered that along with the adoption of digital data in the past 20 years, they have had to confront compatibility issues among operating systems, applications and document formats. Solutions have been to develop new object-oriented programs that wrap legacy Fortran and C code, or to develop entirely new programs that consist of interchangeable modular components, both strategies based on application interoperability (e.g., Microsoft's Component Object Model – COM -- and Sun Microsystems' Common Object Request Broker Architecture). Generally, most new development seeks to maintain links with at least its immediate precursor. This has led to the continued support for legacy formatting of input and output.

Like SHEF, XML is not a database, but an architectural framework for data commonality. Much in the manner that COM defines a general organizational structure, while allowing for individual customization, XML seeks to establish the same governing structure for data. However, unlike the COM model, XML standards are open-source. Furthermore, the World Wide Web Consortium, in accordance with its mission to lead web development "to its full potential by developing common protocols that promote its evolution and ensure its interoperability,"[6] oversees modifications and enhancements to the XML standard. This standard has been adopted by computer industry leaders such as rivals Sun Microsystems and Microsoft. Almost all web-based applications include embedded XML support as a basic component of their software development kits on all computer platforms in wide use today. In addition, XML support is embedded in most common desktop applications and web browsers.

Recall that in the early days of computing, a standard input file consisted of a card deck with its limitations of 80 characters per line, and position based categorization. Changing data requirements necessitated both modifications of the code and of the proscribed data format. XML overcomes these and many other restrictions by defining a general standard analogous to an outline (Figure 2).

- **I.** Hydrologic time series language
  - **A.** Regular-interval time series
    - **1.** Location
      - **a.** location
      - **b.** watershed
      - **c.** station
    - **2.** Parmeter
      - **a.** parameter
      - **b.** parmtype
      - **c.** units
    - **3.** Values
      - data
      - data
    - .
    - .
    - .

Figure 2. Outline representation of XML-based hydromet schema

For each specific use, the XML standard allows a common “dictionary”; in essence an appendage to the global requirements that can be modified as necessary. Moreover, any user can append additional structure without affecting either a specific user community, or the global standard. Finally, even if the underlying standard changes, old files are not made obsolete, but can be utilized by adding or modifying one of several components.

What is XML?

XML “can be used to store any kind of structured information, and to enclose or encapsulate information in order to pass it between different computing systems which would otherwise be unable to communicate.”[7] Figure 3 shows an example of SHEF type E data downloaded from the California Data Exchange Center (CDEC). The same data is shown in Figure 4 using an XML scheme. For most of use not weaned on SHEF, the XML is much easy to decipher because it is self-documenting. Furthermore, the XML is machine-readable.

```
:  Hourly FLOW, RIVER DISCHARGE (CFS) data from CDEC:
:  Station SACRAMENTO RIVER AT I STREET BRIDGE (IST), sensor 20, from
:  01/02/1997
:  00:00 to 01/02/1997 01:15
:  Data subject to revision
:
.E IST 20030409 P DY1997010200/QRH/DIH01 90440/91400/92600/93800/95040
: End of Data
:  Hourly RIVER STAGE (FEET) data from CDEC:
:  Station SACRAMENTO RIVER AT I STREET BRIDGE (IST), sensor 1, from
:  01/02/1997
:  00:00 to 01/02/1997 01:15
:  Data subject to revision
:
.E IST 20030409 P DY1997010200/HGH/DIH01 26.11/26.35/26.65/26.95/27.26
: End of Data
```

Figure 3. SHEF Type E

```
<htsl>
  <rts>
    <loc location="SACRAMENTO RIVER AT I STREET BRIDGE"
     station="IST"/>
    <parm parameter="FLOW" parmtype="PER-AVER" duration="" version=""
     units="CFS" shefpe="QR"/>
    <footnote>sensor 20</footnote>
    <structure initial="1997-01-02T00:00:00" interval="P1H"/>
    <values>
        <d>90440</d>
        <d>91400</d>
        <d>92600</d>
        <d>93800</d>
        <d>95040</d>
    </values>
  </rts>
  <rts>
    <loc location=" SACRAMENTO RIVER AT I STREET BRIDGE"
     station="IST"/>
    <parm parameter="STAGE" parmtype="PER-AVER" duration="" version=""
     units="FT" shefpe="HG"/>
    <footnote>sensor 1</footnote>
    <structure initial="1997-01-02T00:00:00" interval="P1H"/>
    <values>
        <d>26.11</d>
        <d>26.35</d>
        <d>26.65</d>
        <d>26.95</d>
        <d>27.26</d>
    </values>
  </rts>
</htsl>
```

Ordered pair define an element

Elements can have either required (parmtype) or optional (units) attributes to further describe data

Figure 4. A hydromet XML prototype (based on McFadden, 2002)

For those used to coding with HTML, the "tag" organizational concept is familiar. Nonetheless, some description is in order. The < *tagname*> and </tagname> identify all the text between as belonging to the basic building block for XML – the element. Note that elements can be nested, so that in this case, the entire figure is an ***htsl*** element (***h****ydrologic* ***t****ime* ***s****eries* ***l****anguage*), a user-defined subset of XML. The ***rts*** (***r****egular-interval* ***t****ime* ***s****eries*) element is a child of ***htsl***. This nesting process is unlimited. In XML, the structure can be defined using either a Document Type Definition (DTD), or a more robust tool, the schema (Figure 5). One major benefit of the schema is that data verification and validation can be embedded as part of the design. Either method offers tremendous flexibility. Need a custom attribute such as the agency responsible for the data location – add it. Want to tie a rating table to a stage data collection location – define a new child element. Both one-to-one (a location name) and one-to-many (time series data) relationships can be defined. In short, the XML standard can be adapted to fit almost any need for data exchange.

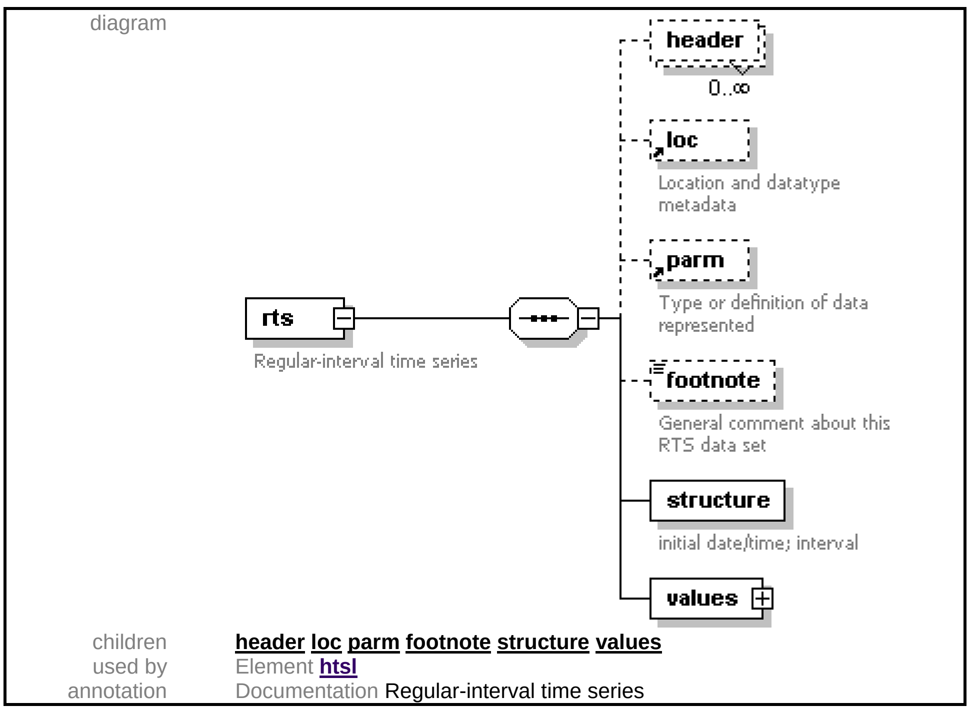

Figure 5. The regular time series element of the XML schema (McFadden, 2002)

How to integrate XML

The process of implementing an XML-based file or data exchange consists of developing translation programs. The work can be divided in half; one-half translates from XML, the other to XML. The first is simply a parsing program built around existing (in most cases) text-based output generators that produce files similar to Figure 4. The second can employ the use of ***e****xtensible* ***s****tylesheet* ***l****anguage* ***t****emplates* (XSLT), another component of XML that provides a standard for rendering XML into a variety of other formats. Because XML is open-source and multi-platform, these programs can be written in most programming languages and take advantage of the common programming architectures in Windows and Sun. The key is defining a hydromet XML that will serve, as in Figure 6, as the center of an exchange wheel. Once implemented, programmers, data processors and data consumers can standardize future development around a common format.

This format is capable of handling not only hydromet data, but can also serve as the basis for exchanging watershed information used in modeling. Underneath the hood of most hydrologic and hydraulic models is basic set of information consisting some combination of measurable physical variables (reach length, watershed area, etc.), parameters that are loosely based on physical characteristics (Manning's roughness, infiltration rates, etc.), and black-box variables (SCS curve number, length to centroid, etc.) that aggregate characteristics and responses. In many cases, this information can be directly exchanged between models, or at the least serve as an adequate initial value. Furthermore, as models are upgraded or revised, the configuration format changes, while the configuration information remains the same.

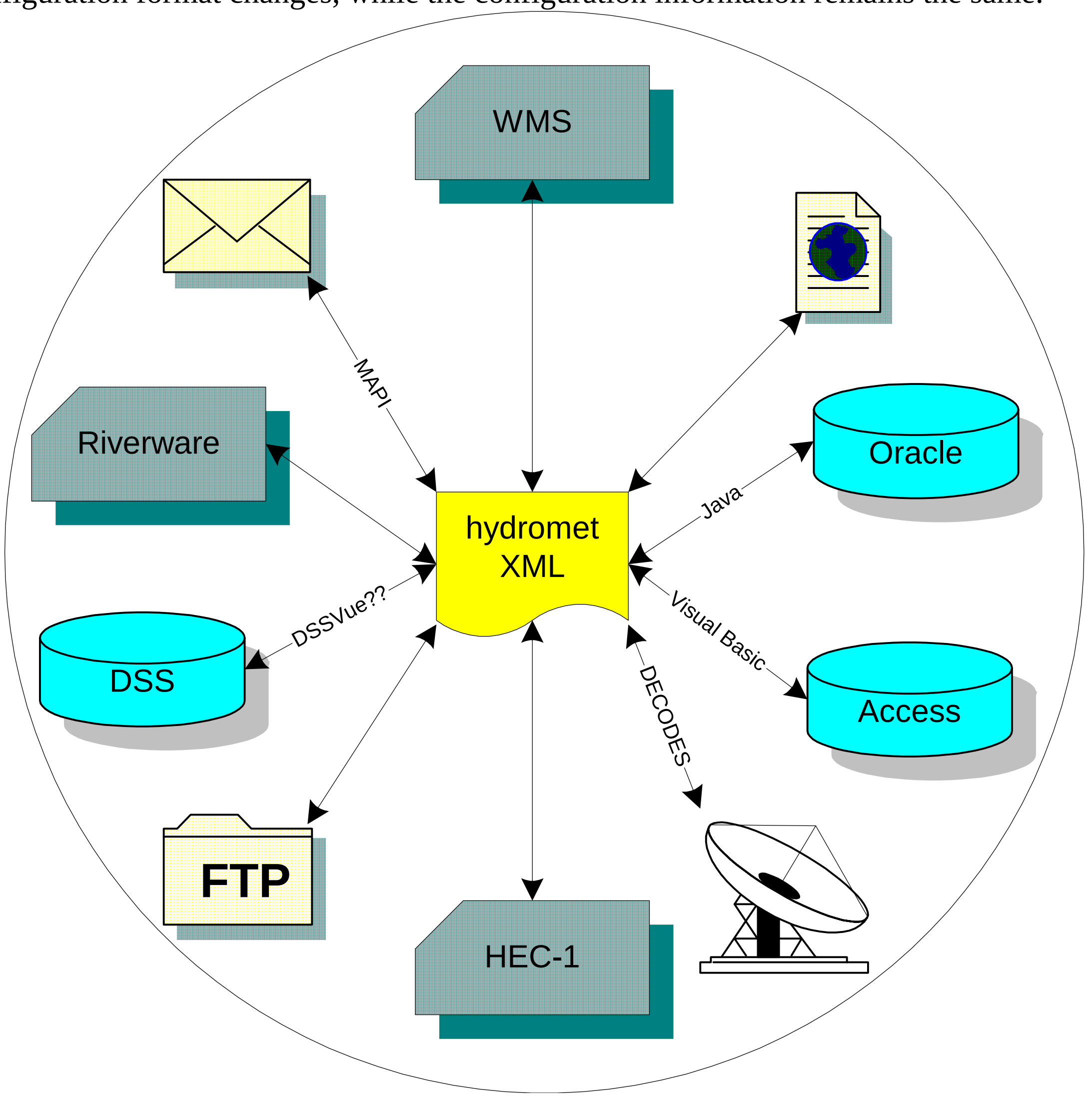

Figure 6. XML centered file exchange wheel

Another benefit of an XML-based standard is security. Computer and data collection networks have become increasingly large. Design and maintenance of a hydromet system that was once undertaken by an individual is now divided into many roles – communications specialists, network engineers, database administrators, data technicians, and hydrologists and engineers. Each role requires limited, but clearly defined responsibilities and access. If your office is like mine, engineers are constantly developing and modifying data query tools, determining how best to manage their data so that they can reuse data in different applications and for different studies. Although most current system configurations prevent unauthorized access to core databases, data incompatibilities still require keyboard based postings, and manually initiated processing. Occasionally, this has created problems with unintentionally deleted data or program components. (Imagine what intentional destruction could do!) Most often, these are correctable through tape backups or other data protection schemes. An XML hub can engender an automated data cycle from input to dissemination through phone, radio, and satellite networks, in essence sequestering data and data processes from inappropriate access. In addition, as web-based data exchanges pervade, this hub provides a common interface for data retrieval. Data acquired in this manner, can undergo customized formatting designed to provide specific users with only *their* essential information, i.e, separating data consumers from data processing for enhanced security.

Conclusion

Most agencies and organizations maintain status quo even in the light of better technologies because of 1) cost; 2) lack of skilled or experienced staff; and 3) bureaucratic inertia. Establishment of a hydromet XML would address the first and second by enabling data management methods to seamlessly integrate with modeling programs. Hydrologists and engineers would dramatically reduce time spent converting data between disparate, incompatible formats. Currents standards were developed for a command line environment and require constant tinkering as data needs evolve. New engineers come with web-based skills, and can readily adapt to an XML-driven environment. Under an XML, future data requirements can be appended to an established flexible format without invalidating files based on an older standard. Costs will be reduced because we no longer will have to develop new software tools to accommodate a new format. More importantly, developers of hydrologic modeling software will quickly recognize the benefit of a commonplace universal data standard and upgrade existing software for XML compatibility. Data collection platform manufacturers such as Sutron have already started implementing XML into newer designs. Databases and web browsers have embedded support for XML.

There is support for XML at the highest levels of government. There is support for XML in the IT community. XML provides tools to better serve stakeholders, partners, and the public. The technology speaks for itself. It time to address number three and start working with other federal agencies, with state and local interests, to embrace and adapt an hydromet XML.

References

[1] http://www.m-w.com/cgi-bin/dictionary?va=extensible
[2] http://www.whitehouse.gov/news/releases/2002/12/20021217-5.html
[3] http://xml.gov/documents/in_progress/strategy.htm
[4] SOFTWARE SYSTEM DESIGN AND IMPLEMENTATION FOR ENVIRONMENTAL MODELING : A MOU WORKING GROUP
[5] SHEF Handbook pg 1
[6] http://www.w3.org/Consortium/
[7] http://www.ucc.ie:8080/cocoon/xmlfaq#acro A-2