Journal of Hydroinformatics | 8.2 | 2006

A generic metadata description for hydrodynamic model data

A. K. M. Saiful Islam and Michael Piasecki

ABSTRACT

Sharing of data sets between numerical models is considered an important and pressing issue in the modeling community, because of (i) the time consumed to convert data sets and (ii) the need to connect different types of numerical codes to better map inter-connectedness of aquatic domains. One of the reasons for the data sharing problem arises from the lack of sufficient description of the data, or lack of metadata, which is due to the absence of a standardized framework for these metadata sets. This paper describes the development of a metadata framework for hydrodynamic data descriptions using the *Geographic Information Metadata, 19115:2003* standard published by the International Standards Organization (ISO). This standard has been chosen not only because of its extent and adequacy to describe geospatial data, but also because of its widespread use and flexibility to extend the coverage. The latter is particularly important, as further extensions of the metadata standard are needed to provide a comprehensive metadata representation of hydrodynamics and their I/O data. In order to enable the community to share and reuse numerical code data sets, however, they need to be published in both human and machine understandable format. One such format is the Web Ontology language (OWL), whose syntax is compliant with the Extensible Markup Language (XML). In this paper, we present an extensive metadata profile using the available elements of ISO 19115:2003 as well as its extension rules. Based on the metadata profile, an explicit specification or ontology for the model data domain has been created using OWL. The use of OWL not only permits flexibility when extending the coverage but also to share data sets as resources across the internet as part of the Semantic Web. We demonstrate the use of the framework using a two-dimensional finite element code and its associated data sets.

Key words | data exchange, geographic information, hydrodynamics, metadata, model, ontology

A. K. M. Saiful Islam
Institute of Water and Flood Management (IWFM),
Bangladesh University of Engineering & Technology,
Dhaka,
Bangladesh
E-mail: *akmsaifulislam@iwfm.buet.ac.bd*

Michael Piasecki (corresponding author)
Department of Civil, Architectural and Environmental Engineering,
Drexel University,
3141 Chestnut Street
Philadelphia,
PA 19104,
USA
Tel: +1 215 895 1721
Fax: +1 215 895 1363
E-mail: *Michael.Piasecki@drexel.edu*

INTRODUCTION

Over the last decade, the accessibility of data through internet-based systems has become more widespread with a plethora of web portals to access real-time data sources, forecasting systems or direct access to holdings in databases. This trend has enabled the hydrodynamic processes modeling community to gain access to a much denser array of data that can be used for calibrating and validating hydrodynamic models, in addition to providing data sets for model improvement. While this access to data is a very welcome development it has also prompted the need to develop a common vocabulary and grammar to exchange model data among users of various models in order to ease the burden of re-formatting numerical model data when using different codes, see for example the Hydrology XML Consortium (HydroXC 2005). A key to the successful development of such a framework is the adequate description of the model data through a metadata system. Metadata is commonly known as "data about data" (Ahmed *et al.*

doi: 10.2166/hydro.2006.017

2001) and is used to provide conceptual information about a data object. Unfortunately, metadata has always been treated as a less important aspect in the world of numerical modeling (Sen 2003). Typically the information about numerical model data is placed in a few lines in the header and throughout the data files, an approach that is common for many of the currently used hydrodynamic models.

Very few hydrodynamic and water quality modeling environments such as BASINS (EPA 2004) have made progress to overcome this deficiency. BASINS provides an integrating modeling environment, which uses metadata to describe the content, quality, condition and other characteristics of model data. Another initiative addressing data interoperability is the *Harmon* IT project (http://www.harmonit.org/) funded by the European Commission that aims at developing and implementing a European Open Modeling Interface and Environment (OpenMI) that will simplify the linking of hydrology related models across Europe (Moore 2004). Other modeling environments such as SMS (ems-I 2004), GenScn (USGS 2004) or DELFT3D (Delft Hydraulics 2001) provide some data interoperability but require the use of specially designed components that are part of a code family or work only with specific set of codes that have been equipped with a custom interface. There are also a number of self-describing data formats available like netCDF (NCAR 2004), HDF5 (NCSA 2004), the HDF-based XMDF (EMRL 2005) or the Standard Hydrometeorological Exchange Standard (SHEF 2005) that is used by the US Geological Survey (USGS) and the National Oceanographic and Atmospheric Administration (NOAA) to exchange data between hydrologic models. In all of these formats metadata is either embedded in the file (netCDF and SHEF), which require those files to be opened to find out what has been stored, or can be stored externally (HDF5), which makes it slightly more attractive for use. However, the meta information stored is not regulated (following a standard) and can be set arbitrarily, leaving it to a user to define what it should be. Even though all of the above formats and approaches have found more or less widespread acceptance, they fail to provide a suitable platform that could be used to interchange hydrodynamic model data. By suitable we mean a data and metadata exchange and description system that is (i) platform-independent and generic so the information can be automatically parsed using standard parsing tools, (ii) uses a metadata standard that is internationally recognized and (iii) that has a higher degree of generalization (less customization of legacy code interfaces) than what is currently available.

METADATA STANDARD AND ITS REPRESENTATION

A number of metadata standards have emerged over the past few years. Among them are The Dublin Core Metadata Initiative (DCMI 2002), the Federal Geographic Data Committee (FGDC 1998) and the International Standards Organization (ISO) recommendation for geographic information metadata, ISO 19115:2003 (ISO 2003). Although the ISO 19115 standard has been primarily developed for digital datasets, its applicability can be extended for use of other forms of geographic data such as maps, charts and textual documents and general-purpose data. Because hydrodynamic model data mostly deals with geographically referenced data sets, this standard is ideally suited for developing a hydrodynamic modeling metadata set.

ISO defines an extensive set of metadata elements (about 300) but typically only a subset of them is used. In addition, while ISO 19115:2003 contains many of the desired description elements, links to other norms within the 19000 group are necessary to deal with specific aspects, like ISO 19108 for temporal schema (ISO 2002), ISO 19103 for units of measurements (ISO 2004) and ISO 19111 for spatial referencing by coordinates (ISO 2000). Although a large extent of standard metadata for digital geographic data has been documented in ISO 19115:2003 (and its associated norms), the standard does not contain all metadata descriptions that are needed to describe data sets of hydrodynamic codes. As a result, the ISO standard needs to be extended, which can be done by creating a community profile for that specific user group. This profile is composed of three parts: core metadata components (23 elements), some elements of the comprehensive metadata set (these are pick-and-choose from the existing definitions) and the extended metadata components (to be defined by the community). The selection of optional metadata in addition to the core metadata is left to the community metadata creators. However, a balance must be found between excessive description of data (the "more-is-better approach") and a sufficient but small enough description set for usability.

The need to pass metadata instances between researchers (human readability) and computer systems (machine readability) requires the selection of a suitable encoding scheme. A first choice would be to use the eXtensible Markup Language (XML) developed by the World Wide Web Consortium (W3C) because it is a simple, very flexible text format that has been designed for electronic publications of any kind (Quin 2003). However, plain XML provides only a syntactic framework, with very little semantic or meaning-between-data-elements capabilities, and also is quite limited in its ability to be easily extended when encoding a metadata profile. Alternatively, one can use the Web Ontology Language (OWL) that is intended to provide a language that can formalize the domain knowledge with explicit specification in a machine readable format (Smith *et al.* 2003). In addition, it provides the advantage that OWL allows a better representation of the ISO 19115 standard in machine readable format and also prepares the metadata profile to be published as a web resource which would permit participation in the Semantic Web (Berners-Lee *et al.* 2001). The OWL language has been applied in a variety of ways in geographic information metadata systems to better address consistency, validation, interoperability needs, verification and required system integrity (Hanschuch & Staab 2003; Wariyapola *et al.* 1999). The availability of the ISO 19115 norm in OWL, even though not (yet) recommended by ISO, is a basic building block for the proposed model data metadata profile as it allows (i) the use of the metadata norm in machine readable format and (ii) the utilization of the inherent characteristics in OWL to express the meaning and relations between elements of numerical modeling data sets.

HYDRODYNAMIC MODEL DATA ONTOLOGY

Typically the desire to exchange input/output data between different codes requires multiple conversion routines to move data from one format specification to another. A better approach to this time-consuming process is to introduce a framework in which data is stored format-independent so it can be parsed (on the accompanying metadata) and re-inserted into other codes without reformatting. Also, in view of the fact that model input and output (I/O) data is either geo-referenced, concerns temporal scales or addresses code-specific operations, it should be categorized to provide some structure. This structure is provided by using a data type and accompanying metadata ontology that permits the description of the associated data and their relations to each other. In addition, an intriguing aspect of this ontology is that it is being made available as a web resource, i.e. it is stored as machine readable structure that can be used as a template by an infinite number of users and applications to generate and parse their data file descriptions. Because the ontology permits the extension to accommodate code-specific description elements the framework has the potential of growing over time to cover an increasing number of code-specific data descriptions, which is in essence the idea of the Semantic Web.

Figure 1 shows the detailed expansion of the ontology top class *IO_Model* that focuses on the actual data sets of the system (on the left) and the metadata description block (on the right). The latter contains the code-specific data set (we use an in-house two-dimensional FEM code as an example that is depicted as *MD_FEM2D*) and the link to the classes describing the spatial and temporal extent of the input and output data (marked by the dotted line rectangles). The code-specific descriptions are relatively small in size and can be added without much effort, allowing the modeling data interchange system to grow in the future by adding more specific modules to the group “others”. It should be noted that the open arrowheads identify properties of a class, while the closed arrowheads depict the relationship of a parent and child class (or sub-class). The difference is that the properties can be range restricted (only a certain set of values can be used when creating instances) and the sub-classes can be given a different set of properties other than those they inherit from the parent class. These differences are quite significant and will not be further elaborated on here, yet the careful identification of sub-classes and properties is a crucial step of the development work as it determines the ability to properly identify and also extend this code framework.

METADATA PROFILE FOR HYDRODYNAMIC MODELING COMMUNITY

We have divided hydrodynamic model data into two basic categories: (i) basic geospatial data and (ii) data related to

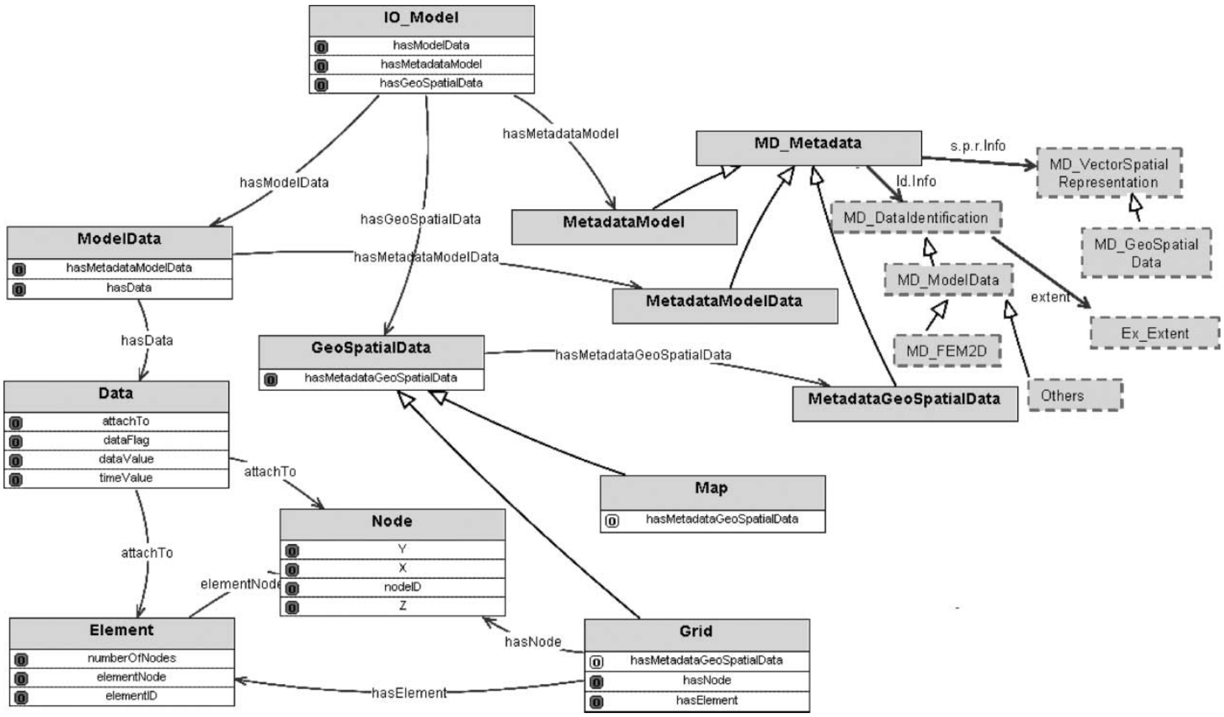

Figure 1 | Data and metadata ontology for hydrodynamic codes.

geospatial data which we call *model data*. Into the basic geospatial data category fall representations of the numerical grid, maps, digital terrain models and boundary polygons. These data are typically time-invariant and could be vector type, such as the grid description, or raster type, such as a topological map, and are necessary to define the geospatial extent of the model domain. The Model data category, while also geospatially referenced, contains the data that is required to describe physical processes (often also time variant) such as wind speed, water velocity, discharge, water level, viscosity coefficient, roughness coefficient, dispersion coefficient, flow direction, boundary types, tidal elevation or tributary discharge data. Besides these two data categories a hydrodynamic model typically also contains parameters that are not geospatially referenced at all, like run time controls, logical flags, fixed parameters (like the acceleration constant) and so on. These are of course quite code-specific because no two numerical codes are alike.

Consequently, we propose two specialization groups of metadata entities to represent hydrodynamic model data: (i) *MD_ModelData* and (ii) *MD_GeospatialData* as shown in Figure 2. The *MD_ModelData* metadata entity was defined as a specialization of the *MD_DataIdentification* metadata entity to uniquely identify resources of the hydrodynamic model. Inside the *MD_ModelData* metadata entity a set of metadata elements was created which describe the data types of the model (e.g. discharge, velocity, dispersion coefficients and so on) and a code-specific class. In this example, we have created a specific class *MD_FEM2D* that is an extension from the *MD_ModelDat*a class and that is unique to the *FEM2D* code. This class contains FEM2D-specific information such as the parameter that controls the use of the Petrov–Galerkin scheme and the number of iterations required, to name just a few. The *MD_GeospatialData* metadata entity was defined as a specialized class of the *MD_VectorSpatial Representation* class to represent vector type geospatial data used in a hydrodynamic model, such as the model grid. Because a model grid consists of a number of nodes and elements the *MD_GeospatialData* description includes the total number of nodes and elements, a code list of permissible element types and the grid type. Also, notice that the data types can be associated with a specific location of the element,

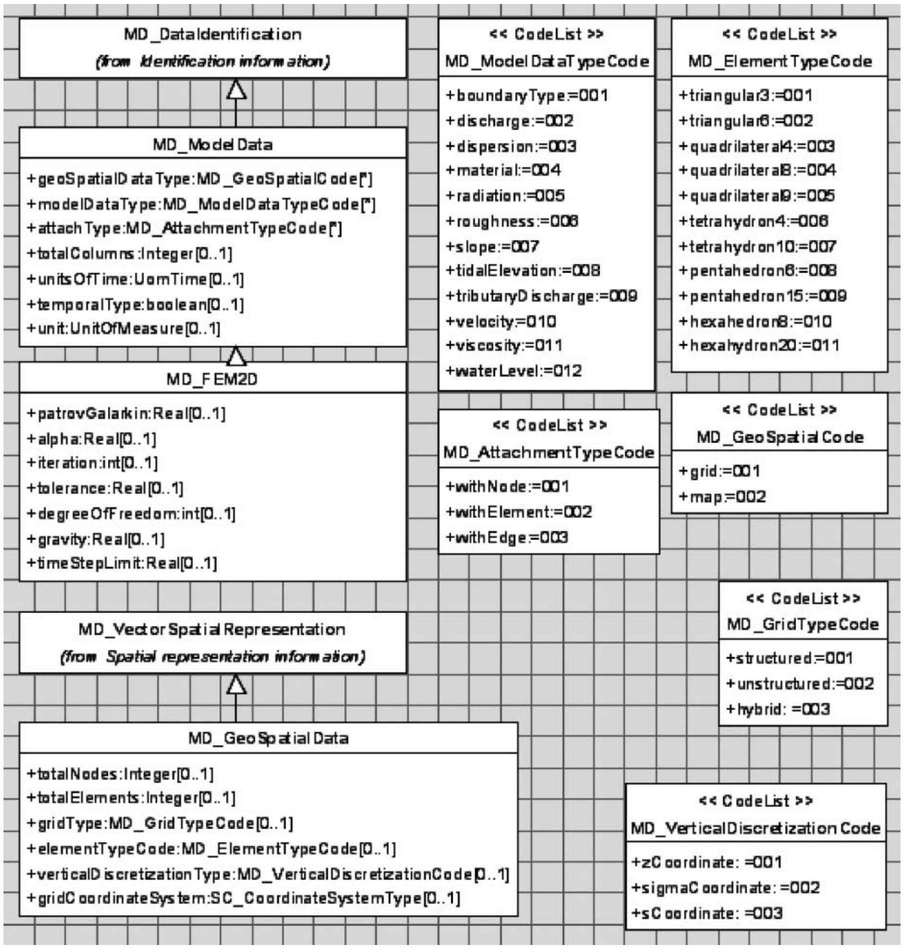

Figure 2 | List of ISO 19115 metadata extensions and code lists for hydrodynamic codes.

i.e. center, node or edge, to reflect the various grid definitions deployed when using FEM, FVM or FDM formulations.

The code list for *MD_ModelDataTypeCode* (Figure 2) is not complete, of course, as many more data types are possible within a numerical code, like pressure, turbulent kinetic energy, energy dissipation, eddy viscosity and eddy diffusivity, to name just a few that, for example, are inherent in three-dimensional code. Our intention here is to provide an example only (for FEM2D), yet we realize that this list must be extended in the future to encompass all possible data types, possibly including a finer granularity (common data types, 1D types only, 2D types only, 3D types only) for better organization and categorization of the metadata.

The number of metadata extensions is quite extensive (44 elements) and covers specific parameters for code operation, element types, grid types, the grid coordinate reference system (horizontal and vertical) and also extensions in the 19103 norm for units of measurement (inclusion of measures like discharge and dispersion). There are also code lists (permissible keywords) that contain different permissible element types for finite element (difference) grids, non-associated geospatial data, an enumeration of different grid types, an open list for different input and output constituents used in the hydrodynamic model, which are attached to the grid, and an open list of vertical discretization types of geospatial data. We have summarized the necessary metadata extensions and codelists in a data directory that, because of space limitations, cannot be shown here. However, we have made these extensions available in tables that can be accessed and viewed at http://loki.cae.drexel.edu/~wbs/data/wbsModel/

AN EXAMPLE DATA/METADATA SET FOR THE FEM2D CODE

As a consequence of using the FEM2D code as an example instance, we require the creation of twelve ModelData file pairs (for each of the 12 *MD_ModelDataTypeCode* entries shown in Figure 2, one for the metadata description, one containing the raw data) and one GeospatialData file pair containing the grid information. These numbers may differ from code instance to code instance as more or fewer constituents are being identified as necessary model data. The exact number is determined in the model-specific sub-class, as mentioned before. Because each of the metadata files contain several hundreds of code lines and the raw data files potentially many thousands, we will restrict ourselves to use only the GeospatialData file as a demonstration. The file contains the grid information (*x*, *y* coordinates and bathymetry, as well as the connectivity matrix) for a model domain (about 23 km in length) encompassing the upper Potomac River adjacent to Washington, DC, as shown in Figure 3, and is comprised of 1171 quadrilateral elements and 1408 nodes.

The complete set of descriptive entries in the grid_metadata file encompasses 23 elements, for which a total of 80 property values must be supplied. Part of the metadata content is shown in the left panel of Figure 4. The top lines contain the Universal Resource Locator (URL) identifier for the raw data file (that can reside anywhere on the WWW) and the lower block refers to information regarding the grid. The right panel shows a segment of the node coordinate descriptions (top) and a subset of the connectivity information (bottom). The formatting of the raw data file in this instance is format-less, i.e. the numbers are only space separated, which allows for easy parsing. The chosen number types are float (coordinates) and integer (connectivity), which are ISO standard types.

In creating metadata instances for each of the previously mentioned data files, it is important to recall that each metadata file will contain the core element set (required by ISO 19115) as was discussed in the second section. This can be done automatically in many instances because the values required repeat themselves in each metadata file. The second step is to select a subset of all other optional and conditional elements that can be related

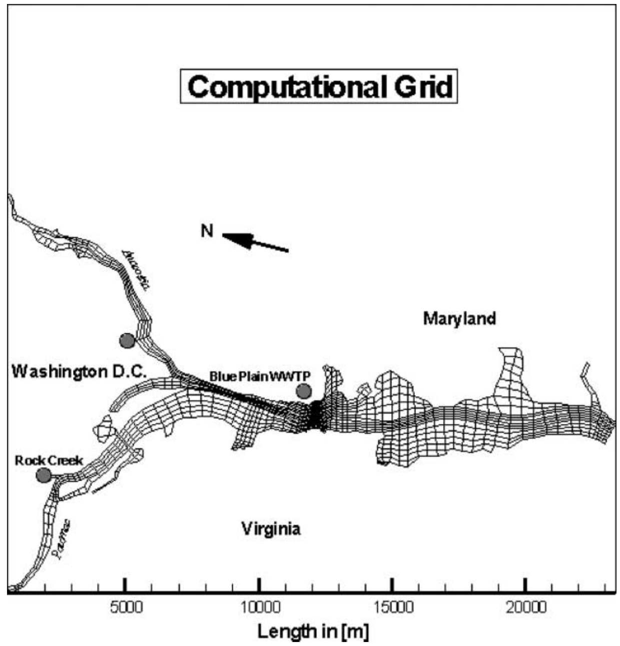

Figure 3 | Numerical grid of the Upper Potomac River.

to the grid metadata description. For example, one can add the e-mail address and telephone number as additional information elements to the *MD_ResponsibleParty* class of this metadata set. The final step is to fill in the extended elements of metadata as described in the third section. The grid files are being made available and can be viewed and downloaded at http://loki.cae.drexel.edu/~wbs/data/wbsModel/, which also hosts a pair of example files containing time variant information associated with the velocity field.

SUMMARY AND FUTURE WORK

In this paper we have proposed a generic metadata description framework for hydrodynamic model data that can potentially pave the way how to interchange and exchange data between different types of codes and model domains. More specifically we have:

- selected the ISO 19115:2003 metadata norm as a basis to generate a generic metadata framework to describe hydrodynamic code data. We have demonstrated the ability of the ISO standard to provide a very generic base set of elements (and properties) for this purpose that is

```
<iso19115:MD_Metadata rdf:ID="Metadata_Gird">
  <iso19115:language>en</iso19115:language>
  <iso19115:characterSet rdf:resource="#utf8" />
  <iso19115:dataSetURI
  rdf:resource="http://loki.cae.drexel.edu/~wbs/
  data/wbsModel/in/grid/data/grid.htm"/>
  .........................
  .........................
</iso19115:MD_Metadata>

<MD_GeoSpatialData rdf:ID="Grid_MD_GeoSpatialData">
   <gridType rdf:resource="#unstructured"/>
  <verticalDiscritizationType rdf:resource="#zCordinate" />
  <gridCoordinateSystem rdf:resource="iso19111:Cartesian" />
   <geoSpatialDataType rdf:resource="#grid"/>
   <elementType rdf:resource="#quadrilateral4"/>
   <totalElements>1171</totalElements>
  <totalNodes>1408</totalNodes>
</MD_GeoSpatialData>
  .........................
  .........................
```

Selection of GRID metadata elements

```
1      697.504     593.720    28.042
2      647.517     694.853    28.042
..............
..............
1407   23367.492   6559.296   15.240
1408   23402.545   6620.561   15.240
```

Sample of node coordinate information

```
1     4  1     4     5     2
2     4  2     5     6     3
.........
.........
1170  4  1395  1406  1407  1396
1171  4  1396  1407  1408  1397
```

Sample of element connectivity

Figure 4 | Samples from the example metadata and grid data files.

based on the set of recommended core elements as well as a number of extensions. These extensions (elements, properties and code lists) are embedded into the ISO 19115 norm where necessary. Also, we have shown how certain aspects of the metadata structure, like time and units, needed a linkage to other ISO metadata standards, such as ISO 19103 (units) and ISO 19108 (time).

- used the Web Ontology Language OWL as an encoding medium for the hydrodynamic metadata profile, because (i) it best captures all aspects of the Unified Modeling Language, UML, in which the ISO norms are presented and (ii) permits the use and access of these schemas on the Semantic Web, i.e. makes it accessible as a resource that can be automatically parsed for content. In other words, all elements, their properties and their relation to and among each other are manifested in an ontology that can be changed and updated as the profile might change.

It is clear that the scope of this framework is extensive and would need to be filled over time. Hence, future work should be directed towards several goals. Firstly, the current specifications need to be expanded to include more code-specific blocks, for example for HEC-RAS, SWMM, HPSF or EFDC, to name just a few. With the growing number of data file descriptors more code-specific data sets can then be parsed by other applications to populate their input data files. Secondly, in order to make this a seamless approach the system would require a toolset that automatically handles the submission of a data file of type X that is parsed, cut into the appropriate pieces, described via the suggested metadata blocks and then send back (encoded in XML or OWL) to the point of submission. Here the owner can either make the files available on a web server (then they become parsable) or keep it on a local hard disk. Another scenario would be to search for any data (let's say velocity) on the web, find a file that has been described using this system and simply download the file (together with the grid) to obtain an instance (or many) of a velocity field at a certain time for a specific simulation scenario. Because the raw data file is well described through the companion metadata file and is also referenced to a grid that itself is referenced via a global standard Coordinate Reference System, a user can identify velocities at specific geographic locations within the grid. Finally, in addition to providing tools that would help to convert legacy data sets in format Y into machine readable data sets that are well described, one should also develop a repository that (i) can be used to archive interfaces that automatically convert the XML/OWL data sets into specific code-dependent formatted data sets in case the user does not want to alter his copy of the code and (ii) that can be used to store data sets for any code in the generic format for accessibility by the WWW.

The basic premise behind this approach is that of using the Semantic Web by making these descriptions accessible as a resource to any user, not only as a document that can be read but also as a resource that can be accessed from web-based applications that use the published resource both for description purposes and for finding and parsing data files. We believe that this approach can aid in overcoming the currently existing lack of data interoperability among different hydrodynamic codes.

ACKNOWLEDGEMENTS

This work has been partially supported by a grant from the National Oceanographic Partnership Program (NOPP), grant number NAG13-0040, and by the National Science Foundation (NSF) under grant number 0412904. This work was conducted using the Protégé resource, which is supported by grant LM007885 from the United States National Library of Medicine.

REFERENCES

Ahmed, K., Ayers, D., Birbeck, M., Cousins, J., Doods, D., Lubell, J., Nic, M., Rivers-Moore, D. & Watt, A. 2001 *Professional XML Metadata*. Profession. Wrox Press, Birmingham.

Berners-Lee, T., Hendler, J. A. & Lassila, O. 2001 The Semantic Web. *Sci. Am.* **284** (5), 34–43.

DCMI 2002 *Dublin Core Metadata Initiative*. Available at: http://dublincore.org/

Delft Hydraulics 2001 *DELFT 3D: Simulation Models and Modelling Systems Related to Integrated Coastal Zone Management*. Available at: http://www.netcoast.nl/tools/rikz/delft3d.htm

EMRL 2005 *XMDF*. Environmental Modeling Research Laboratory, Brigham Young University. Available at: http://emrl.byu.edu/xmdf1/index.html

ems-I 2004 *Surface Water Modeling System (SMS)*. Available at: http://www.ems-i.com/SMS/SMS_Overview/sms_overview.html

EPA 2004 *Better Assessment Science Integration Point and NonPoint Sources (BASIN)*. Available at: http://www.epa.gov/docs/ostwater/BASINS/

FGDC 1998 *Content Standard for Digital Geospatial Metadata (CSDGM)*. Available at: http://www.fgdc.gov/metadata/contstan.html

Hanschuch, S. & Staab, S. 2003 CREAM: creating metadata for the Semantic Web. *Comput. Net.* **42**, 579–598.

HydroXC 2005 *The Hydrology XML Consortium*. Available at: http://www.nws.noaa.gov/oh/hydroxc

ISO 2000 *Geographic Information – Spatial Referencing by Coordinate System (ISO 19111)*.

ISO 2002 *Geographic Information – Temporal Schema (ISO 19108:2002)*.

ISO 2003 *Geographic Information – Metadata (ISO 19115:2003)*.

ISO 2004 *Geographic Information – Conceptual Schema Language (ISO 19103)*.

Moore, R. 2004 *The OpenMI Architecture – Report A: Scope and organisation – Version 0.9*. Available at: http://www.harmonit.org/documents/pubview.php?view = all&doctype = all

NCAR 2004 *Network Common Data Form (netCDF)*. Available at: http://my.unidata.ucar.edu/content/software/netcdf/index.html

NCSA 2004 *Information, Support, and Software from the Hierarchical Data Format (HDF) Group of NCSA*. Available at: http://hdf.ncsa.uiuc.edu/

Quin, L. 2003 *Extensible Markup Language (XML)*. Available at: http://www.w3.org/XML/

Sen, A. 2003 Metadata management: past, present and future. *Decision Support Systems* **37** (1), 1–23.

SHEF 2005 *Standard Hydrometeorological Exchange Format*. Available at: http://nws.noaa.gov/om/whfs/SHEF/Shef_2.pdf

Smith, M. K., Welty, C. & McGuinness, D. L. 2003 *OWL Web Ontology Langauge Guide*. Available at: http://www.w3.org/TR/2004/REC-owl-guide-20040210/

USGS 2004 *Generation and Analysis of Model Simulation Scenarios (GenScn)*. Available at: http://water.usgs.gov/software/genscn.html

Wariyapola, P. C. H., Abrams, S. L., Robinson, A. R., Streitlien, K., Partikalakis, N. M., Elisseeeff, P. & Schmidt, H. 1999 Ontology and Metadata Creation for the Poseidon Distributed Coastal Zone Management System, *IEEE Advances in Digital Libraries Conference, ADL '99*. IEEE. New York. Availabe at: http://citeseer.ist.psu.edu/wariyapola99ontology.html