An XML Schema for Digital Communication Signal Research

Kenneth Holladay
Kay Robbins
Denise Varner

Abstract

A priori analysis and classification of digital communication signals is important both in the communication industry for developing new receiver technology and in the intelligence community for deciphering intercepted signals. In both domains, algorithm research focuses on extracting signal characteristics, such as modulation type and symbol rate, even when the received signal is distorted by noise and fading in the transmission channel. The lack of common standards for test files, data recording, and presentation of results make it difficult to compare the performance of algorithms from disparate sources. This work proposes an XML-based framework and an associated XML Schema as a solution to this standards problem. The XML Schema defines base types for communication signal constructs. These base types are then used to define two principal XML document formats: a signal library and an algorithm test record.

The XML signal library provides a standard for cataloging and exchanging test signal files, a need that has been acknowledged for some time. We have implemented modules in Matlab and Simulink that generate signal files in WAV format while recording the signal attributes in the XML signal library format. XSLT scripts provide an easy way to peruse the library using a standard web browser. The results of our work validate many of the XML claims for standard data representation. However, we also discovered that developing a good XML Schema is a non-trivial undertaking that requires proficiency in both XML and the application domain. Those just adopting XML should be prepared for a considerable time investment to master Schemas and XSL.

Table of Contents

1. Introduction ........................................................................................................ 2
2. Signal Classification .......................................................................................... 2
   2.1. International Standards ........................................................................... 3
   2.2. Ad Hoc Amateur Radio Schemes .............................................................. 3
   2.3. Proprietary Schemes ................................................................................ 4
3. XML Schema Development ............................................................................... 4
   3.1. Single Carrier Signals ............................................................................. 4
   3.2. Multiple Carrier Signals ......................................................................... 5
   3.3. Multiple Carrier Signals that Vary in Time ............................................. 6
   3.4. The Digital Communication XML Schema ............................................ 7
4. Extending the XML Schema to a Signal Library .............................................. 7
   4.1. Building a Library ................................................................................... 7
   4.2. Using the Library .................................................................................... 8
5. Discussion ......................................................................................................... 9
6. Appendix - The digicom.xsd Schema ............................................................. 10
   6.1. Carrier Description Type ....................................................................... 11
   6.2. Segment Description Type ..................................................................... 11
   6.3. Signal Description Type ........................................................................ 12
Bibliography ......................................................................................................... 12
1. Introduction

Many disciplines are rapidly adopting XML as the universal format for information exchange. With its siren song of simple structure, published standards, platform independence, and ubiquitous support, XML’s attraction can conceal the practical challenges of developing a truly usable XML Schema. Successful adoption within a technical discipline requires expert knowledge of both the application domain and XML technology, as well as political savvy in forging agreements with potential adopters.

This paper chronicles the evolution of an XML Schema designed to facilitate research in the area of a priori digital communication signal analysis, particularly in the High Frequency (HF) communication band. The HF band is defined as the frequency range from 3 to 30 MHz. In practice, most HF radios use the spectrum from 1.6 to 30 MHz. In this range, the ever-present hazards of noise, fading, and interference make establishing and maintaining a viable HF communication link more difficult than in the VHF (very high frequency) and UHF (ultra high frequency) bands. However, HF signals (especially in the range of 4 to 18 MHz) have the unique ability to bounce off of the ionosphere, enabling them to move information around the world instead of just across a line-of-sight [HC96]. Government regulations and international treaties divide the band into sets of frequency ranges for specific communication purposes including maritime, aviation, distress, standard time, amateur, broadcasting, and radio astronomy [NTIA].

The development of satellite communications and the proliferation of VHF and UHF radio repeaters resulted in a declining interest in HF communication for many years. That trend is now reversing. Satellite communication channels are crowded and their cost is increasing. Also, recent technology advances have overcome many of the former problems associated with HF communication, thus renewing interest in HF as a viable and cost-effective long-range communication medium [KE02].

While analog signals (e.g., broadcast voice and music, and amateur radio) constitute a significant percentage of the traffic on the HF band, most of the research and development activity is in digital communication. A priori monitoring, analysis, and classification of these digital signals is important both in the communication industry for developing new receiver technology and in the intelligence community for deciphering intercepted signals. Current emphasis is shifting from manual monitoring techniques that require a skilled operator to computer controlled automated processes.

As with any technical discipline, agreed-upon standards for information interchange facilitate the accurate and timely dissemination of data among interested parties. For automated computer processing, an information exchange standard must also be syntactically rigorous, semantically unambiguous, and technically complete.

The remainder of the paper is organized as follows. The next section presents an overview of some existing techniques for classifying HF signals. We then describe the development of the basic XML Schema for signal specification and how it was used to build a catalog for a test signal library generated using Matlab and Simulink [MATH]. The paper concludes with a brief discussion of the process used to forge agreement among the parties that are adopting the XML Schema. The result, digicom.xsd, is presented pictorially in the Appendix. The files referenced in this paper, including the current version of the signal XML Schema and an example test signal library, are available online at ftp://ftp.swri.org/pub/signals/2002/.

2. Signal Classification

The basic theory for digital communication is well established, and there are many good books available [PR00] [SK01]. Digital information is encoded in a transmitted signal by varying in time one (or more) of three fundamental characteristics: frequency, phase, and amplitude. This process yields a variety of forms ranging from the simple carrier amplitude modulation of Morse code to intricate schemes such as Voice Frequency Telegraphy (VFT) that exploit all three types of modulation. Table 1 lists common abbreviations associated with frequently encountered basic modulation techniques.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW</td>
<td>Carrier Wave. This is the same as OOK (On Off Keying) or Morse code.</td>
</tr>
</tbody>
</table>
FSK | Frequency Shift Keying. This is the same as PFM (Pulse Frequency Modulation), also known as PSM (Pulse Skipping Modulation)
---|---
MSK | Minimum Shift Keying
PSK | Phase Shift Keying. This is the same as PPM (Pulse Phase Modulation)
DPSK | Differential Phase Shift Keying
OQPSK | Offset Quadrature Phase Shift Keying
ASK | Amplitude Shift Keying. This is the same as PAM (Pulse Amplitude Modulation)
QAM | Quadrature Amplitude Modulation. This is the same as APSK (Amplitude and Phase Shift Keying)
CPFSK | Continuous Phase Frequency Shift Keying
CPM | Continuous Phase Modulation
GMSK | Gaussian Minimum Shift Keying

Table 1. Common Digital Modulation Classifications

The acronyms of Table 1 provide a general classification of signal types, but they do not completely specify a signal. There are numerous other parameters associated with each modulation strategy, and some of these techniques can be combined in a single broadcast signal. Most modulation strategies can be described by analytical formulas. However, this form is not suitable for a general signal library specification, because real signals have noise and are distorted by channel fading and other effects. A complete specification must include all of these parameters in a flexible manner to accommodate signals that have been recorded off the air as well as generated signals. Even when incompletely specified, real signals are useful in generating test suites for exercising classification algorithms.

2.1. International Standards

The International Telecommunication Union (ITU), headquartered in Geneva, Switzerland is an international organization within the United Nations System where governments and the private sector coordinate global telecom networks and services. They publish a Radio Regulations book that contains a detailed emissions classification scheme [ITUR]. The intent of this scheme is to identify emission sources for regulatory and compliance monitoring.

The ITU scheme consists of a four-character expression specifying signal bandwidth, followed by a five-character encoded description of the emission. For example, 2K11H2BFN represents a “selective calling signal using sequential single frequency code, single-sideband full carrier with a bandwidth of 2.11KHz.” Selecting the appropriate letters and numbers is sufficiently complicated that the International Amateur Radio Union has published a simplified guide that reduces the 15-page standard down to a 2-page table with 15 selections that cover the most common signal types [IARU]. They state that in ambiguous cases, anything else can be classified as “Unknown.” The ITU standard may serve its intended purpose and could be encoded in an XML Schema, but it does not contain sufficient detail to describe an arbitrary digital signal.

2.2. Ad Hoc Amateur Radio Schemes

Amateur Radio enthusiasts constitute another active group of people listening to signals in the HF bands. Companies such as Monteria [MO02], Klingensfuss [HC96], and others publish lists of frequencies, descriptions, and sample recordings of monitored signals. The Worldwide Utility News organization publishes a widely distributed “Frequently Asked Questions” (FAQ) document that describes and classifies many of the signals heard on HF [SC97]. Since the target audience for this information consists of people who are listening to radios, the principal classes in this FAQ are distinguished by how the signals sound. For example, “Synchronous Data Block” signals are described as having a “distinctive chirping sound,” while “Synchronous Bit Stream” signals are “continuous and possess a trilling quality.”
Within each of the principal classes is amassed a significant quantity of additional technical information. However, the presentation of this information is not without problems. Some terms are used inconsistently, while others can be quite ambiguous. For example, the term “baud” (the number of symbols transmitted per second) and the term “bps” (the number of bits transmitted per second) are sometimes used interchangeably. Also, ambiguous words like “tone” are used to describe both a symbol value in a simple FSK modulation, as well as one of many subcarriers in more complicated modulation schemes. Once again, this description format is not well suited for computer based analysis and information exchange.

2.3. Proprietary Schemes

Finally, several government organizations and commercial companies have designed and built automated signal acquisition and analysis systems with varying degrees of capability [SWRI]. Many of these systems use databases to store signal descriptions, but there are no published standards for tables or field names.

While a significant amount of academic research has been published on algorithms for signal classification, there is no standard format for test files, data recording, or presentation of results. This lack of common standards makes it difficult to compare the performance of algorithms from different sources.

3. XML Schema Development

Clearly, signal analysts and researchers would benefit from a more rigorous signal description mechanism, like an XML Schema, that addresses the shortcomings of current schemes. Ideally, the XML Schema should define domain specific data types and elements that can unambiguously describe almost any digitally encoded signal while retaining a structure that can support typical automated operations such as cataloging, searching, and sorting.

Designing an XML Schema of this magnitude is not a trivial undertaking. There are many critical decisions such as key names, attributes, and hierarchy levels. In most cases, there are multiple ways to achieve the same objective, with no clear indication of which is better. We chose to decompose the problem, beginning with simple signals then progressing to more complicated signals. This section presents an abbreviated chronicle of that process. Representative signal spectrograms are included to help visualize the problem. Spectrograms are a widely used method for visualizing the frequency components of a signal as a function of time. The brightness in the plot area indicates the energy (amplitude) of the signal. In these spectrograms, white represents the highest energy level, and black represents the lowest. The frequency to which the receiver is tuned becomes the zero point on the histogram and is called the baseband signal.

3.1. Single Carrier Signals

Figure 1 shows a spectrogram for a simple modulation type: Binary Phase Shift Keying (PSK with 2 phase states). The spectrogram shows a clear, single energy band 1000 Hz above the tuning frequency. The information in this signal is encoded in phase changes, which do not show up clearly on the image, but can be extracted in other ways. Leaving out a few details, this signal might be described with the following XML snippet.

```xml
<carrier>
  <freq_Hz>1000</freq_Hz>
  <modulation>PSK</modulation>
  <symbolsPerSec>100</symbolsPerSec>
  <numStates>2</numStates>
</carrier>
```

Many of the terms in this construct are specific to signal processing, and for this paper, their particular meaning is less important than the context in which they appear.
Examining signals that use other modulation schemes from Table 1 yields several additional necessary elements. For example, an FSK signal might be described as follows:

```
<carrier>
  <freq_Hz>0</freq_Hz>
  <modulation>FSK</modulation>
  <symbolsPerSec>100</symbolsPerSec>
  <numStates>2</numStates>
  <modIndex>3</modIndex>
  <phaseContinuity>discontinuous</phaseContinuity>
</carrier>
```

We could have defined separate XML complex types for each modulation class, but this introduces unnecessary complication. Therefore, the “carrier” element is defined in the XML Schema as a “carrierDescriptionType” with one required element, the frequency offset, followed by a series of optional elements. See Figure 4 in the Appendix for the complete definition.

### 3.2. Multiple Carrier Signals

Next, consider the class of signals comprised of more than one carrier. Figure 2 shows a VFT spectrogram that has a single unmodulated carrier (the solid white line at about 200 Hz) along with multiple other carriers that are independently modulated. There are two obvious choices for describing this signal: either introduce a new complex type to describe multiple carriers, or add a hierarchy level that allows for multiple “carrier” elements. While the latter approach adds a layer of complexity for simple signals, it has a significant benefit for later processing in that all signal types have the same structure. The result became the “SegmentDescriptionType” illustrated in Appendix Figure 5.
3.3. Multiple Carrier Signals that Vary in Time

Many real signals transmit data in bursts. Although the signal energy turns on and off as a function of time, as long as the modulation characteristics do not change from one burst to the next, the previous “SegmentDescription-Type” provides an adequate definition. However, the “segment” does not allow for the possibility that the signal will change its character midstream. Figure 3 shows a CODAN signal from a smart modem that can monitor the effectiveness of an HF transmission and adapt its modulation to achieve a high data throughput with minimal error rate. The modulation characteristics are different in the two areas of signal energy. To describe this, the segment must capture the signal start and stop time, and an additional hierarchy level is required to contain multiple segments. This is illustrated in Appendix Figure 6.

The XML description for the FSK signal previously illustrated now looks like this.
3.4. The Digital Communication XML Schema

This process of refinement was essential to creating digicom.xsd, a viable XML Schema for Digital Communication Signal Research. Wherever practical, we defined enumerations and restricted data types, so that the XML validation process supports semantically correct use of the types. In addition to the signal description, we defined types to describe the communication channel (noise and fading), as well as types related to saving a signal in a file.

The above XML snippet also illustrates another important design decision. Many elements, for example frequency and time, can be expressed in many different units of measure. Some XML Schemas that we examined provided a “units” attribute. We decided to adopt a fixed unit for each field and to embed the unit as part of the field name. This decision favors ease of searching for values over ease of constructing user displays.

Usually, we preferred elements to attributes. We did define two attributes for general use: a confidence and a tolerance. This provides a straightforward way for an algorithm to report confidence levels for calculations and for specifications to associate tolerances with values.

4. Extending the XML Schema to a Signal Library

The first application of the XML Schema was the construction of a library of test signal files for signal recognition algorithm testing. The problem of obtaining test signal files for this purpose has been recognized for some time. Kremer and Shiels [KR97] stated, “There also appear to be no existing databases of radio transmissions which could be used for our purpose.” In related areas of research, such as speech recognition, significant effort has gone into establishing standard corpora of test data files [UP02]. While there are a few communication signal resources available, such as [DE02] [KL02] [LA01] [RU96], these files usually have insufficient knowledge of one or more important signal attributes to be useful for algorithm evaluation. Also, these repositories have no formal structure and are unsuitable for automated selection of files that meet specific criteria. As a result, most researchers are compelled to produce their own body of simulated signals. Besides the tremendous duplication of effort, rarely are sufficient details of the simulation process available to allow independent analysis of the test data sets used to validate an algorithm.

Due to the relatively large nature of these signal files and to the anticipated quantity of files required, it did not appear to be practical to store the actual signal data in an XML format. Rather, we chose to create an XML catalog for the files. We created a new XML Schema, siglib.xsd, which includes digicom.xsd, thus extending the namespace.

4.1. Building a Library

We selected Matlab and Simulink from the MathWorks as our tools for creating a test signal library. These programs have extensive support for digital signals, and the new version 6.5 includes XML DOM support. The resulting program automatically records each new signal in the library as it is produced. We wrote small support functions...
to create XML fragments that could be easily reused. For example, the following function creates an FSK modulated carrier element.

```plaintext
function carrierNode = mkFSKcarrierNode( docNode, rCarrierFreqHz,... rSymbolsPerSec, nStates, rModIndex, strPhaseCont )
    carrierNode = docNode.createElement('carrier');
    elemNode = docNode.createElement('freq_Hz');
        elemNode.appendChild( docNode.createTextNode( num2str(rCarrierFreqHz) ) );
    carrierNode.appendChild( elemNode );
    elemNode = docNode.createElement('modulation');
        elemNode.appendChild( docNode.createTextNode( 'FSK' ) );
    carrierNode.appendChild( elemNode );
    elemNode = docNode.createElement('symbolsPerSec');
        elemNode.appendChild( docNode.createTextNode( num2str(rSymbolsPerSec) ) );
    carrierNode.appendChild( elemNode );
    elemNode = docNode.createElement('numStates');
        elemNode.appendChild( docNode.createTextNode( num2str(nStates) ) );
    carrierNode.appendChild( elemNode );
    elemNode = docNode.createElement('modIndex');
        elemNode.appendChild( docNode.createTextNode( num2str(rModIndex) ) );
    carrierNode.appendChild( elemNode );
    elemNode = docNode.createElement('phaseContinuity');
    if( 'c' == strPhaseCont || 'C' == strPhaseCont )
        elemNode.appendChild( docNode.createTextNode( 'continuous' ) );
    else
        elemNode.appendChild( docNode.createTextNode( 'discontinuous' ) );
end;
    carrierNode.appendChild( elemNode );
```

4.2. Using the Library

With the XML Schemas in place, and having generated a signal library description file (siglib.xml), our attention turned to XSL as the obvious tool for manipulating the XML data. After carefully studying several good books [HO02] and numerous web articles, our first attempts at XSL scripts produced no output!

Following standard practice [WA02], we had defined a namespace for our XML Schema and made it the default namespace in the siglib.xml file. However, we failed to define the namespace in the XSL file, and we did not explicitly reference that namespace when using “select” and “match” to specify elements. This simple error took several hours to resolve, since there are few good debugging tools for XSL scripts.

We are now in the process of constructing a variety of scripts in support of our research. XSL is proving to be a powerful tool that performs significant work with just a few lines of code, but getting those few lines correct is sometimes a challenge. The concepts of XSL programming are significantly different than most popular programming languages like C, BASIC, C++, or Java. Experienced programmers often have difficulty adapting. If this problem is not adequately addressed, it may prove to be a significant barrier to widespread adoption of XSL [JA02].

As a simple example, the following code fragments demonstrate a way to peruse a signal library using a standard browser. First, we have an HTML file that the browser loads.

```html
<html>
<head>
<script>
    var doc, xslDoc;
    function onload()
    {
        doc = new ActiveXObject("msxml2.domdocument");
        doc.async = false;
        doc.load("siglib.xml");
        xslDoc = new ActiveXObject("msxml2.domdocument");
        xslDoc.async = false;
    }
</script>
</head>
<body>
</body>
</html>
```
Here are the XSL templates that produce the document body.

```
xsl:template match="/"
  <xsl:apply-templates/>
</xsl:template>

<xsl:template match="dc:signalLibrary">
  <h1><xsl:value-of select="dc:libName"/></h1>
  <h3>Number of files: <xsl:value-of select="count(//dc:signalFile)"/></h3>
  <table style="font-family:verdana;font-size:10pt;" rules="all">
    <tr>
      <th>FileName</th>
      <th>SampleRate</th>
      <th>CommonName</th>
      <th>SymbolsPerSec</th>
    </tr>
    <xsl:apply-templates select="dc:signalFile"/>
  </table>
</xsl:template>

<xsl:template match="dc:signalFile">
  <tr>
    <td><xsl:value-of select="dc:fileDescription/dc:fileName"/></td>
    <td><xsl:value-of select="dc:fileDescription/dc:sampleRate_Hz"/></td>
    <td><xsl:value-of select="dc:signalDescription/dc:commonName"/></td>
    <td><xsl:value-of select="dc:signalDescription/dc:segment[1]/dc:carrier[1]/dc:symbolsPerSec"/></td>
  </tr>
</xsl:template>
```

5. Discussion

The results of our work validate many of the XML claims. Our data storage is well organized, and the XML technologies are now mature enough to fulfill our automated processing requirements. Yet the road to adopting XML was filled with difficult decision points. We began seriously considering XML for signal descriptions in December 2001 and developed our XML Schema over the next nine months. The two alternate technologies under consideration involved using a standard database format, or defining yet another proprietary file format.

There are considerable advantages associated with employing a database. Using SQL to define the tables, fields, and relations yields a degree of platform independence, and most database implementations have rich tool sets for manipulating and extracting data. The disadvantages appear when you select a specific database for implementation. The platform independence and easy portability disappears, and our principal simulation tool, Matlab, did not have any native database support. Of course, Matlab did not have XML support at that time either.

Our previous work in this area made liberal use of user defined formats for both binary and ASCII data files. These are easy to build and modify, readily supported by Matlab, and our programmers are familiar with the technology. The principal disadvantage is long-term maintenance. Old records of tests in a comma-delimited format are not very useful if no one remembers which fields appear in what order. Careful naming of XML elements imparts a degree of built-in documentation, while an associated XML Schema provides a perfect repository for detailed descriptions of the elements, their relationships, and any assumptions or restrictions associated with the data.
Finally, the time and effort expended in this project demonstrates that developing a good XML Schema is a non-trivial undertaking that requires significant proficiency in both XML and the application domain. Those just adopting XML should be prepared for a considerable time investment to master XML Schemas and XSL.

6. Appendix - The digicom.xsd Schema

The XML Schema illustrations in this section were produced using XML Spy version 4.3 [AL02] and represent a small selection of the defined types. The full schema is available at ftp://ftp.swri.org/pub/signals/2002/.
6.1. Carrier Description Type

Figure 4. Carrier Description Type

6.2. Segment Description Type
6.3. Signal Description Type

Figure 6. Signal Description Type

Bibliography


**Biography**

**Kenneth Holladay**  
Manager, Signal Exploitation Software Section  
Southwest Research Institute  
San Antonio

XML 2002 Proceedings by deepX
Kenneth Holladay is the manager of the Signal Exploitation Software Section at Southwest Research Institute in San Antonio, Texas. His principal areas of expertise encompass software engineering, software design and development, distributed computing, and industrial instrumentation. Mr. Holladay holds a B.S. in Chemical Engineering from the University of Florida and is a registered Professional Engineer. He is currently pursuing a Ph.D. in Computer Science at the University of Texas at San Antonio.

Kay Robbins
Professor
University of Texas at San Antonio
Department of Computer Science
San Antonio
United States of America

Dr. Kay Robbins is a Professor at the University of Texas at San Antonio in the Department of Computer Science. Her research focuses on visualization, analysis and management of multimedia data sets, particularly those generated from scientific experiments and empirical studies of computer systems. Her work emphasizes time-critical visualization and the comparison of multimedia data sets. She collaborates with experimental scientists in the life and physical sciences to use visualization tools on real data. She holds a PhD in Mathematics from the Massachusetts Institute of Technology.

Denise Varner
Staff Scientist
Southwest Research Institute
San Antonio
United States of America

Dr. Denise Varner is a Staff Scientist in the Signal Acquisition and Radiolocation Department at Southwest Research Institute in San Antonio, Texas. She specializes in formal modeling and simulation of signal recognition systems and applications of pattern recognition and image processing methods. She holds a PhD in Psychology from the University of Pennsylvania.