

Successful Electronic Submission of SGML-encoded Full-text Documents to OSTI

**Julianna U. Hearn
Westinghouse Savannah River Company**

**A paper prepared for presentation at
INFORUM '98
May 6 & 7, 1998**

In early 1993, OSTI identified and published goals for electronic submission of SGML-encoded full-text documents from contractor sites by the year 2000. In response, the SRS STI Program developed a plan for achieving that goal, with an additional goal of publishing those full-text documents to the world wide web. These two goals were incorporated into the site's 1997 Annual Operating Plan (AOP) as Milestones (achievement of AOP Milestones are the basis for determining Award Fees for WSRC). The electronic submission of SGML-encoded full-text documents to OSTI was to be accomplished by July 31, 1997, and publishing of full-text STI documents to the web was to be accomplished by August 31, 1997. This paper explains how these two goals were successfully accomplished ahead of schedule, examines lessons learned in the process, and explores future initiatives.

Developing the Plan

Gathering Information

Before we could begin developing a plan we needed to know a little more about SGML other than how to spell it! Therefore, the first step became locate and attend SGML training. Based on knowledge gained from the training, the following ideas were incorporated into our plan. The decision was made to utilize the DTDs already developed by OSTI. Although rather large and somewhat difficult to decipher, this meant one less thing to worry about creating. A second decision was to keep things as simple as possible for the initial submission by focusing on relatively small, uncomplicated journal articles and conference papers. These efforts will eventually expand to more complicated documents that include graphic, equations, tables, etc. In addition, we determined a need to identify what SGML products and solutions were already available, and what other companies were doing who had successful SGML solutions in place. To this end, we obtained evaluation copies of products from several companies including ArborText and Electronic Book Technologies (EBT). We viewed a demonstration of a system in development at ORNL utilizing Omnimark with an ORACLE database, and we talked with a company in Atlanta that was installing an Intergraph-based system. We also read white papers and attended several seminars sponsored by vendors at which some of their customers explained their successful implementation of SGML document systems.

Identifying Possible Solutions

The Ideal Solution

After reviewing our needs and reviewing what we had learned evaluating products and benchmarking other companies, we developed the consummate solution. The first piece is a “black box” conversion tool for generating HTML- and SGML-tagged documents from word processing documents. Additionally, a database (preferably relational) for storing the documents once they are converted to SGML-tagged. Ultimately, the database interface would be web-based and would include the means for searching full-text, SGML tags, and bibliographic information. Further more, the database would include the means for tracking documents through the STI review and approval process. The programming support assigned to our project assured us such a system could be developed and we eagerly anticipated the realization of our ideal solution. Unfortunately, resources had to be reallocated in the middle of the project and “replacement” support from elsewhere on site could not be obtained due to a site-wide reduction in force.

Alternative Solutions

Fortunately, we had the forethought to identify some alternative solutions. One such solution included utilizing Adobe’s FrameMaker + SGML as the conversion utility for SGML-tagged document generation. Frame + SGML appears to be a viable solution as a conversion tool, but problems importing the OSTI DTD into the program to develop the required rules file negate its usage at this point in time. Another possible solution identified was to use the SGML Author add-on to MicroSoft Word. Since the SRS site-licensed word processing software is MicroSoft Word, and we have the possibility for having thousands of authors submit documents using that software, this appeared to be a good alternative. However, the SGML Author add-on has been abandoned by MicroSoft. It is not offered as an add-on for new versions and is no longer supported for previous versions. Our third alternative solution was to use OSTI’s idea of submitting HTML 3.2 encoded files as an interim option. However, when we discussed this option, it was agreed that we would continue to pursue the true SGML-encoded document solution.

Making It Happen Anyway -- Actual Implementation

In order to meet the AOP Milestone deadlines that were fast approaching, we had to implement the only remaining alternative – hand-tagging the SGML-encoded document (here’s where that SGML training comes in handy!). Even when focusing on small, uncomplicated journal articles and conference papers, this is an enormous and time-consuming task. To help make it more manageable, the task was broken down into several steps. First, a likely candidate document that had already been approved for release through the STI process was identified, and a clean, electronic version in MS Word was obtained from the originator. Then, an HTML version of this document was created using an rtf-to-html conversion tool. Creating the HTML document first helped with identifying the structure of the document and identifying where some of the SGML tags might likely be placed. Next, the OSTI DTD was modified slightly to more closely

represent SRS document contents, and a “template” file was developed for use as a guide when hand-tagging the actual text. Once the SGML-tagging was completed, the document was submitted to OSTI as an attachment to an e-mail message. OSTI parsed the document and returned error listings, as well as suggestions for fixing problems. The document was modified and returned to OSTI for parsing. This process continued until a document that would parse was created and it was accepted by OSTI. In this manner we were able to successfully accomplish both AOP Milestone deadlines ahead of schedule. The first successful submission of a SGML-encoded full-text document to OSTI was accomplished on July 23, 1997. The HTML document required a small amount of cleanup before it could be placed on the web. Then, this same document was published to the external web page on August 7, 1997, after generating the parent HTML page from which it was linked.

Current Process

A process similar to the one for the initial electronic submission of an SGML-encoded, full-text document to OSTI is currently being followed for subsequent submissions. The documents being selected for tagging and submission are still relatively small journal articles and conference papers, but we are getting progressively bolder with more complicated documents that include graphics and tables. A copy of nsgmls has been obtained and is being used to parse the files and correct problems prior to submitting the documents to OSTI. In addition, the submissions to OSTI are being made in the required format of zipped files that include the SGML-encoded full-text document, the related SGML-encoded bibliographic file, and any related graphics files. To date, multiple successful submissions of SGML-encoded full-text documents have been made to OSTI.

Lessons Learned

The following is a list of key tasks or items that we identified as imperatives for successful implementation of electronic submission of SGML-encoded full-text documents to OSTI.

- Know SGML and its rules. Specifically, understand the hierarchy, what is allowed where, what is required, attribute usage.
- Know your DTD. If you create your own, make sure it will generate a document instance (tagged document) that will parse against the DTD OSTI is using. Make sure your DTD itself will parse before you use it. If you decide to use the DTDs developed by OSTI, make sure you obtain the latest versions, and make sure you have all of them (the ostirep.dtd and all its children, the 133215.dtd).
- Make sure the document instance contains the opening statement expected by OSTI’s parser. Initially, I was informed that statement for reports was: `<!DOCTYPE PUBLIC "-//DOE OSTI//DTD OSTIREP//EN" "ostirep.dtd">`, but have since been told that statement is not required by the OSTI parser (a DOCTYPE statement is, however, required by the nsgmls parser). The bibliographic document instance does

not need a DOCTYPE statement in order for OSTI to successfully parse it as long as the tags are correct and in the right place.

- Use the same tag names in the document instance that are in the DTD, make sure they are spelled correctly as per the DTD. OSTI has made the “bluebook” for bibfile submittal available on-line at <http://www.osti.gov/html/osti/eei/bluebook.html> and it is very helpful. However, some of the tags in the bluebook file are not exactly the same as those in the actual 133215.dtd.
- Utilize some sort of SGML parser to check document instances prior to submitting them to OSTI. This parser should be at least as “strict” as the parser used by OSTI. Ideally, use an SMGL-enabled authoring tool to generate document instances or convert word processing documents.
- Know and use the required file naming conventions for electronically submitting files to OSTI. This information is available at <http://www.osti.gov/html/osti/eei/sendel.html>.
- Have some sort of zip or tar utility to properly pack the SGML files being submitted electronically (e.g., WinZip or PKZip).
- Knowledgeable technical support is a must. This includes someone at OSTI willing to work with you and guide you through problems, mistakes, and questions, as well as other knowledgeable SGML consultant types. Quick response from the technical support personnel is a necessity.
- Try small, simple files first. Plan ahead for a phase-in of other product types. Have alternate plans for accomplishing goals. Be flexible!
- Ensure management understands process and the resources it will take – management buy-in is required! Keep them informed of progress and problems.

Looking Forward to the Future

The STIP teams have developed and agreed to the following goals related to electronic dissemination of STI:

- site submission of the announcement record (metadata or bibliographic information) to OSTI in either SGML, XML, or HTML encoded format
- electronic full-text submission to OSTI from sites in SGML, XML, or HTML encoded format, with accepted interim formats PDF, ASCII, MS Word version 5.0 or greater, WordPerfect version 5.0 or greater, TIFF Group 4.
- site submission of the above information via FTP, on electronic media, or URL address.

Once these proposals are incorporated into the Order and Guide, the entire DOE complex will be responsible for developing their own approaches for accomplishing electronic delivery or posting of STI by the year 2004.

SRS will continue to identify and evaluate products. One product currently being evaluated is FileMaker Pro, version 4.0. This version of FileMaker provides the capability to construct a web-based interface to a FileMaker Pro database (including a search tool without writing CGI scripts). This program is being considered for web production of the

SRS STI abstract collection. We are also gathering information from several different companies on “complete-system” products to provide SGML conversion capability and interface to a database for storage of full-text, SGML-tagged documents.

In the near future, we will continue to submit the STI full-text documents as hand-tagged files. Once the new order and guide are published and things become more stabilized, we will evaluate other options, such as a combination of SGML, XML and HTML for STI metadata (bibliographic) and full-text submission to OSTI and web publication. XML is fairly new but appears to offer more flexibility as a means to deliver SGML-encoded information to the web. This could prove to be very beneficial considering the long-term STIP goal of a virtual library based on a decentralized delivery and posting system for STI documents.

In conclusion, we were able to successfully accomplish our goals of web publication and electronic submission of full-text SGML-encoded documents to OSTI because we had a backup plan in place and were flexible enough to put it to use. It wasn't easy, being the first to use new technology or incorporate new methods rarely is. However, we believe information management of STI documents utilizing SGML is the right direction to go. Now, we just have to continue to figure out how to get there!