Trip Report

SGML '92: the quiet revolution

C. M. Sperberg-McQueen
TEI ED R2

December 10, 1992

Version 2, December 10, 1992

The SGML '92 conference, sponsored as always by the Graphic Communications Association and held in Danvers, Massachusetts, was attended by over 275 people, a new high for this conference, and provided good opportunities for learning about SGML or keeping current on what is going on in software and SGML use. Like those of its predecessors I have been able to attend, it owed a lot to the energy and intellectual curiosity of its organizer, Yuri Rubinsky, and was one of the most exciting conferences I have recently attended.

Rubinsky began the conference by passing the year in review, reporting on a bewildering variety of activities. HyTime has been approved as an international standard, the SGML five-year review is in progress, work continues on the Conformance Testing Initiative and the development of SGML-aware query languages (on which more below), and the Document Semantics and Style Specification Language (DSSSL) should come up for a second ballot in early 1993. User groups are being founded left and right, major public initiatives are underway in the aircraft industry (ATA/AIA 100 — I don't swear to the total accuracy of all these acronyms and numbers!), the Commission of the European Community (TIDE, a project using SGML to handle services to the disabled and other persons with special needs), the Unix industry (the Davenport group — Davenport turns out to mean nothing at all, it's just a name they liked — has created a Standard Open Formal Architecture for Browsing Electronic Documents [SOFABED]), the (legal) drug industry, and elsewhere. And of course SGML continues to penetrate the wysiwyg word-processor market.

By the time the Year in Review was finished, the conference was ten or fifteen minutes behind schedule, which persisted as a chronic condition, more to the amusement than to the annoyance of the attendees.

The keynote address was delivered this year by Charles Goldfarb, the father of SGML, under the title “I Have Seen the Future of SGML and It Is ...” He began by reminding the meeting that despite its successes, SGML is not entrenched and has no guarantee of long life: in the larger scheme of current data processing and information technology, SGML is still just a minor blip. He identified several dangers facing SGML in these perilous times. First, the industry continues to view data representation as a minor matter, and to define new data representations for new technologies so as to minimize the effort of using those technologies. The SGML goal of putting the information owner first, and of ensuring that one's data will survive one's computer system, is easily lost in the hustle to design new data representations for new hardware and software systems, as can be seen in the monthly procession of new standards for hypertext and multimedia encoding. SGML apologists must continually articulate the advantages of SGML and of taking the non-obvious approach of suiting the representation to the information, and not to specific hardware devices with a relatively short lifespan. This is not easy in the face of the technology-specific alternatives. Let us face it: it's easier to buy a bunch of Windows applications which can exchange data in the manner peculiar to Windows than to press vendors to support exchange using more rational device-independent systems, which (being device-independent) don't exploit the peculiarities of Windows.

No viable alternative to SGML exists, but competition continues to come in two forms: vendor-promoted technology-specific interchange formats, and turnkey systems which claim to handle all the details. “Let us make all the decisions for you,” say vendors. He noted in particular the mirage of a standard scripting language for multimedia systems, and predicted that it would be the PL/I or the Esperanto of hypermedia: widely heard of and seldom used.

Moving to his main theme, Goldfarb proclaimed the death of the “document”, which he said may in fact never have been anything more than a makeshift to enable the use of computer technology. The future of SGML lies in its use to link both within and between documents. The future of SGML, that is, is HyTime. He showed medieval pages (from the Winchester Bible) and discussed the division of labor among scribes, rubricators, illuminators, and applicators of gold leaf, which corresponds closely to the division of labor, in presenting a hypermedia document today, among the text displayer, the graphics presentation software, and other specialized modules. Hypertext schemes today differ from the methods of the past only in incorporating time-based information. The data structure must be highly optimized to make possible real-time presentation of time-based data, but logically speaking, all that is required are mechanisms for establishing (specifying) synchrony among events. SGML provides a firm basis for representing the abstract information structures needed.

The morning concluded with the first of several poster sessions, which at SGML conferences most resemble high school science fairs. Several speakers were stationed around a meeting room, with wall space for displaying posters on which they had summarized their presentation, and chairs in front of them for auditors. The audience had ninety minutes to move from one to another of the posters, and periodically the chair of the session wandered through the rooms ringing a set of bells as a reminder to the auditors to move to other posters, and to the speakers to begin again from the beginning for the new auditors. Apart from occasioning a rash of jokes about pastoral beasts, the bell system was felt to work very well, and when one of the later poster-session organizers omitted the bells, there was a general request that they be restored.

As a presenter in this session, I was unable to get to any of the interesting posters, and so missed presentations on the creation of modular DTDs, the use of parameter entities in DTD maintenance, and a method for using Post-It Notes in DTD design (saves crossing out). I spoke about the Pizza Model of DTD construction used by the TEI.

After lunch, Susan Hockey and Don Walker gave an overview of the Text Encoding Initiative, describing its organization with its attendant advantages and disadvantages, and focusing on the intellectual problems posed by the broad, varied user community, the internationality of the user community and the project, and the use of volunteers in development of a DTD.

Peter Flynn followed with a description of the CURIA (Cork University and Royal Irish Academy) project to make machine-readable encodings of extant Irish texts in all languages, from the sixth to the sixteenth centuries. He compared the project to similar corpus projects, outlined its projected uses in lexicography, literary research, historiography, hagiography, political science, and folklore. The texts will be in SGML, using the ISO 646 Internal Reference Version character set, and TEI-conformant as far as possible; they will be made available by anonymous ftp, by telnet to the textbase, through the World-Wide Web, on CD-ROM, and by interactive messages to a server. The DTD includes provision for marking titles, authors, names of places and persons, events, dates, numbers, occupations, and shifts of language. He also described some of the particular problems posed for name marking by adjectival prefixing and discontinuous cardinal numbers in Irish. He capped the presentation by remarking that for obvious reasons the tags used would all be in Latin, and providing a Latin expansion for the acronym SGML: Stantis Generalis Monstrationis Lingua (which means: Standard Generalized Markup Language).

The most exciting paper of the day, for me, was George Kerscher and Yuri Rubinsky's paper on "SGML and Braille, Large Print and Voice-Synthesized Text: Work of the International Committee for Accessible Document Design." Kerscher, who for several years ran a non-profit organization called Computerized Books for the Blind and Print Disabled, is now Director of Research and Development for Recording for the Blind, and chair of ICADD. ICADD is seeking ways of making current international standards like SGML and ODA bear fruit in making texts more accessible to print-disabled readers (ten million in the U.S. alone); the flexibility in output styling provided by well designed SGML applications means a text can be presented on a refreshable Braille screen, in a character-based format readable by standard voice synthesizers, in large print, or in other forms, to suit the requirements and preferences of the reader. The structural information provided by SGML is also extremely useful in making it possible to produce grade-2 Braille from machine-readable texts, since Braille symbol usage depends heavily upon context and genre. To exploit the promise of SGML, ICADD is defining a set of architectural forms providing the distinctions most useful in machine generation of Braille, and encouraging developers of other DTDs to provide mappings from their elements to the ICADD architectural forms. Yuri Rubinsky offered to send full documentation to DTD developers, and received a small flood of business cards.

The afternoon was filled out by reports from the standards front. ISO 9070, providing for registration of SGML public text, is moving toward implementation. ANSI was originally named to serve as the registry but wishes to transfer this responsibility to the GCA, which will be happy to do it. The GCA Conformance Testing Initiative is moving forward, but needs money; this led to a spirited discussion of whether formal conformance testing was a Good Thing (all hands up), whether it was a Necessary Thing (almost all hands), and who wanted to try to persuade their management to help pay the quarter to half million dollars needed to complete a serious test suite (two or three hands). No one seems to care whether Turbo Pascal is ISO-conformant or not (it isn't), so I wondered why so many people wanted third-party certification of SGML processors, but there were a lot of government suppliers present, and they explained that procurement rules can make certification attractive or even absolutely necessary. Anders Berglund of ISO reported on the Harmonized SGML Math Initiative, which is effecting a merger of the tags for math in ISO TR 9573-1988, the AAP DTD, and the Euromath project results. (I was surprised to learn that the Euromath project had produced a tag set oriented to the typographical layout of the formula on the page, rather than the logically or semantically oriented markup I had expected — one that would allow arithmetic expressions, for example, to be imported from SGML into spreadsheets or computer algebra programs; the difficulty of providing full semantic markup for all of known mathematics appears to have deterred them from attempting such a scheme.) Further discussions of math markup were held during the week, but I was unable to attend. Finally, Sharon Adler reported on the status of DSSSL, DIS 10179. DIS version 1 was passed in August 1991, but the work group elected to revise the standard further. Version 2 is expected to go out for ballot in April 1993. DSSSL works on the SGML document tree, not on the SGML data stream, using a declarative language to describe processing and a computational component to enable arithmetic computation of some attribute values.

The evening of the first day was occupied by a Novice's Guide to HyTime, which I would have liked to attend, but missed. Reports were that the handout was very useful, so I got a copy of that.

The later days of the conference, though equally full, left less distinct impressions on me. The second day began with a panel organized by Tommie Usdin, who had asked five SGML professionals to design DTDs for the *New Yorker*, giving them however different design goals. Debbie LaPeyre designed a DTD to conform as far as possible to the AAP DTD; Dennis O'Connor designed a DTD to produce the typography of the magazine; Halley Ahearn to load the material into a retrieval system; Yuri Rubinsky to capture as much as possible of the semantic content of the magazine (using what many attendees called "content tagging" to my initial mystification); and Steve DeRose, who worked with David Durand, to produce a hypertext-oriented DTD. The differences and similarities of the DTDs were extremely interesting, as were the different styles of presentation and documentation.

The poster session on the second day was devoted to vendor demonstrations, with demos by vendors of:

- retrieval systems, including Open Text Systems (full-text databases)
- SGML editors and publishing systems, including CAPS/Agfa (high-end publishing), DataLogics (SGML Writer Station), Frame Builder (structured wysiwyg word processing), Arbortext (ditto), Interleaf (showing Interleaf 5 SGML), and Xerox (showing DocuBuild, which "does all the things all the other guys' stuff does")
- application development tools, including SoftQuad (demoing an Application Builder program which enables deep customization of Author/Editor and provides an object-oriented version of Scheme as a programming language) and Software Exoterica (demoing OmniMark, an SGML-aware programming language suitable for data conversion and other processing)
- conversion tools and services, including U.S.Lynx (conversion services), Zandar (demoing TagWrite, a data conversion tool), TMS Inc. (services), Avalanche Development (demoing Fast-Tag), and Data Conversion Laboratory
- others, including Silicon Graphics (reporting on their experiences putting all their online documentation into SGML), George Kerscher demonstrating adaptive equipment, and showings of *SGML: The Movie* (which I once again failed to see)

The third day saw a series of presentations on DTD development by the Society of Automotive Engineers (working on SAE J2008, a DTD for automotive service manuals, maintenance advisories, etc.), the Air Transport Association / Aerospace Industries Association Rev. 100 (ditto for airplanes), and the Davenport Group (including the Committee for the Common Man [Page]). All the speakers were good, but Diane Kennedy's presentation on ATA/AIA Rev 100 was outstandingly clear and factual. Notable in the Davenport presentation was their quick adoption of HyTime architectural forms in the Davenport Advisory Standard for Hypermedia (DASH). A poster session devoted exclusively to problems of tables frustrated many people, who wished it were possible to hear the problems discussed at greater length than the ten or fifteen minutes possible in the poster session. I heard Anders Berglund speaking about the deficiencies of current table markup standards for producing tables of moderate complexity as exhibited by several examples of ISO tables, and Bob Barlow giving a tutorial on the CALS table tags. Both made me glad that other people are working on these problems and that the TEI can simply use their results.

In the afternoon, after a number of case studies, came a long series of talks on SGML query languages, which provided some of the intellectual high points of the conference. Tim Bray of Open Text Systems gave a clear and cogent presentation on "SGML as Foundation for a Post-Relational Database Model." He drew disturbing analogies between current text processing methods and general data processing methods of the period before consistent database modeling and database use:

- files belong to applications
- it's a good application if it produces nice printout
- data sharing only by conversion to different formats
- ad hoc access? forget it
- intolerable application backlog

He suggested that MIS "saved itself" by consistent use of data modeling systems, database access / data manipulation languages, indexing, 4GLs and GUIs, and providing administrative features like

concurrency control, transaction support, audit trails, etc., all crucially linked with the relational data model. He proposed further that text processing save itself the same way: by using SGML as a data modeling language, developing SGML-aware data manipulation and access languages, using indexes for performance, and so on, but emphatically not using the relational model as basis, since it has such a very poor fit with textual data. Given the recent brouhaha on comp.text.sgml over the use of SGML for data modeling, I was struck by the remark "I believe strongly that SGML is a very good language and system for modeling text databases in the real world." I gather that in Waterloo there is more variation of opinion than I knew.

Bob Barlow and Fritz Eberle then described an SGML view of databases, using a somewhat more detailed image of how such a database can be put together and how it works. I was startled, though, to hear that "Editing does not go on inside the document management system; this is a repository."

Paula Angerstein described the background for a panel on SGML query languages which took the rest of the day, with time out for dinner. An SGML query is, she explained, merely a question about what is in an SGML document — a means for identifying interesting pieces of an SGML document, usually for retrieval and possibly for processing. The panelists had each been given a list of thirteen queries to perform on a sample text, or at least to formulate. For example,

- locate all paragraphs in the introduction of a section that is in a chapter that has no introduction
- locate all sections with a title that has "is SGML" in it
- locate all topics referenced by a cross-reference anywhere in the report

(The full list and the sets of solutions ought to be posted separately, as an interesting set of queries and answers.)

After Angerstein's presentation, the members of her panel each spoke briefly about the languages in question. Francois Chahuneau spoke about the language SGML/Search, which he has defined for use in a variety of projects and implemented on top of the PAT indexing engine from Open Text Systems. In SGML/Search, the first two sample queries given above may be expressed:

```
<para> within.1 <intro> within.1 <section> within.1 <chapter> no
containing.1 <intro>
<section> containing <title> containing "is SGML"
```

The third query is not expressible in SGML/Search, which has nothing resembling the required implicit join. It also does not treat SGML attributes as distinct entities and thus cannot formulate the queries involving attribute values.

Paul Grosso described the DSSSL query grammar in general terms: it works on the document tree, using relationships between nodes as defined by a preorder traversal of the tree, with both structure and content accessible to the query. Queries can begin from the root node, or from any set of objects in the document tree, and all queries have the same syntax:

```
query(input-object-set, relationship, criteria, subset)
```

A special notation is provided for queries beginning at the root, using the relationship DESCENDANT.

Steve DeRose described the HyQ query language defined by the recently adopted HyTime standard. The basic data object of HyQ is the *node-list*, an ordered list of locations. The *nodes* in the list may be in one document or several, on one machine or spread across the world, and may be at any level from a single character (or conceivably lower, as in a bit-mapped image of letters on a page) to the world. HyQ

provides a number of low-level functions accepting node-list arguments and returning Boolean or node-list values. Like SGML/Search, HyQ provides normal set operations of intersection, union, difference, etc. (even though node-lists are not necessarily free of duplicates, as sets are).

Neil Shapiro concluded the session by describing SFQL (Structured Full-text Query Language), a query language based on SQL and developed by the Air Transport Association for client/server applications, which extends SQL by adding concepts of fields, proximity searching, fuzzy matching, retrieval control via relevance-ranking etc., and extended data types.

The evening session, attended mostly by die-hard SGML enthusiasts and techies, proved a mixed experience. Shapiro offended many attendees with a pedantic and occasionally patronizing description of SQL (repeatedly implying, for example, that standard SQL is incapable of string searches in text fields, which is not true in the documentation of standard SQL which I have seen), and the audience responded with increasingly rude questions and increasingly questionable claims about the problems inherent in applying the relational model to textual data. Dubious exaggerations to the effect that the relational model must inherently lose information present in an SGML-tagged text, and utter irrelevancies like objections to SFQL on the grounds that its indexing must take more storage than that of other query languages, led in turn to acid requests from yet further members of the group (yes, including me) that the discussion return to substantive issues and attend to technical, not political, issues.

In the confusion, neither the strengths of the sample SFQL implementation of the sample data and sample queries, nor the very real conceptual weaknesses of the relational view of structured text, received adequate attention. After the first stormy hour or so, the discussion did become more substantive and slightly less tense, and eventually it became clear that all of the languages were in fact capable of formulating most of the queries. Francois Chahuneau and Tim Bray provided models of technical objectivity and equanimity as they pointed out forthrightly where the formalisms of their query languages were unable to express some of the desired queries. (PAT and SGML/Search do not, for example, handle the query “Locate all sections with a title that has 'is SGML' in it, allowing the string to be interrupted by sub-elements”.) The others, being unconstrained by existing implementations, were able to demonstrate syntax that allowed the formulation of the queries, without needing to demonstrate engines that can handle those formulations; this slight irony did not pass unobserved by the audience, though no one ventured to point it out publicly.

The final morning was devoted to a series of talks on product creation, including what were reported to be very good talks by John McFadden on the creation of MicroSoft Cinemania (an SGML-encoded hypertext movie encyclopedia) and Ken Kershner of Silicon Graphics on the conversion of paper documentation to a CD-ROM. (I couldn't attend these talks, being busy timing my own talk in my room.)

Eric Freese's discussion of data conversion at Mead Data Central (also praised by those who heard it) led into a final poster session devoted to problems of data conversion.

The closing keynote address was delivered in the dining room over dessert and coffee, which gave it an engagingly relaxed atmosphere. In it, I attempted to match Charles Goldfarb's account of SGML's future with my own predictions of the problems that will occupy the SGML community in the coming years: the development of a fuller consensus on what constitutes good style in DTD design, the need for application portability (not just portability of data), and above all the need for better understanding of the semantics of SGML documents and their processing. In attempting to come closer to useful semantic specifications of SGML DTDs and application processes, six topics should be explored: specialized DTDs for DTD documentation, synonymic relationships among tags (e.g. “**<bold>** is synonymous with **<hilited rendition = 'bold'>**”), class relationships among element types, the operations allowed and forbidden to act upon given element types, axiomatic semantics, and reduction semantics. The looks on the audience's faces ranged from beaming smiles to interested attention to slightly apprehensive puzzlement (especially during the discussion of first-order predicate calculus).

At the conclusion of the conference, the audience gave Yuri Rubinsky, the organizer, a loud and well deserved ovation.