

Creating electronic texts and images

Electronic Text Centre, University of New Brunswick Libraries

Peter D. James University Archivist

University of Winnipeg 2003 CIDL Bursary Recipient

The Electronic Text Centre at the Harriet Irving Library is a Canadian leader in the field of electronic texts and images providing education and leadership to the New Brunswick and Canadian communities in the creation and use of electronic text and electronic scholarly publishing.

We came from all over North America for the Centre's seventh annual Summer Institute. We were university librarians from Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, and Alberta; faculty and educationalists from schools and government; and people engaged in the field of digital creation in Canada and the United States.

We had come to build skills in the world of electronic texts under the direction of David Seaman, Director of the Digital Library Federation of the United States and former director of the University of Virginia Electronic Text Centre.

The UNB Centre's Director, Alan Burk welcomed us; and, Lisa Charlton, the Centre's leader in SGML/XML Initiatives, was instrumental in setting up the weeklong workshop, providing tours of the Text Centre, and facilitating our learning experience.

The Institute got going early the first morning with introductions, and David Seaman began the instruction with a description of the various standards that are at the heart of the production of electronic text: SGML, its offspring XML, and HTML. The course was premised on the use of XML as the current standard descriptive tagging instructions for electronic texts. He pointed to a number of advantages of XML: simple and flexible, it loses little of SGML's descriptive abilities while being more web-friendly, and its inter-operability increases distribution across systems. To employ the jargon: it is cross-platform, software and hardware independent. In short, XML separates description from appearance, relying on style sheets (XSL) that can be read by web browsers for presentation.

... importance of using document type definitions ...

Seaman spoke of the importance of using document type definitions when marking-up documents, even though XML does not require a formal DTD. The value of a formal DTD, like the Text Encoding Initiative (TEI), is that it validates the tags and promotes conformity with universal standards.

On day two, Seaman concentrated on the presentation of web material. The session included discussion of optimum versus practical capture and output of images and text, Optical Character Recognition, and metadata. He discussed project planning with an eye to costs and inputs.

... practical experience in the field was of real value ...

Seaman's practical experience in the field was of real value in explaining the choices project planners must address: consider the use of out-source data providers, engage them to also do the 'dirty' OCR work; factor in the use and pricing of equipment like scanners and digital cameras for projects, and related topics. He mentioned the advantages and disadvantages of several commercial capture and 'clean-up' packages. For project managers, he discussed issues like descriptive standards, rights management, contract staff versus employees, budgeting and related subjects.

In this area, as in all his instruction, Seaman tempered his enthusiasm for his mission with practical advice and knowledgeable tips. We viewed exceptional websites, along with modest ones. For those who will begin to include digital and electronic text creation in their portfolios, it was gratifying to learn that many of the hesitations, second-guessing, and related agonizing that go into projects in 'new' areas, have been addressed by practitioners, and there is plenty of useful (and usable) instructional material on the web (like W3 Schools).

... 'consistency is more important than accuracy' ...

The classroom points Seaman made were reinforced when the group was invited to the Text Centre's Digital Imaging Centre for a demonstration of image capture and manipulation by the staff. The Imaging Centre provides material both for the Electronic Text Centre and for outside contracts. The scale of their operations justified expenditures on equipment and staff, many of whom were trained at the UNB campus.

The bulk of the course was devoted to a hands-on mark-up workshop of

correspondence files that make up part of the University's Charles G.D. Roberts fonds. The Text Centre will eventually produce an electronic collection of the correspondence of this Canadian writer and man of letters, and Institute participants have now contributed (however expertly) to that project. David Seaman introduced the group to the TEI, a DTD employed in the humanities.

The recurring mantra of this three-day workshop: 'consistency is more important than accuracy' (try telling that to your cataloguing colleagues) was offered by Seaman as both an invitation and a caution: part of XML's ease of use is that the tags are not pre-defined, as in HTML, but this is also part of the difficulties one can encounter when constructing an XML tag set.

A tag set can be well-formed (that is, it can possess correct syntax) but not valid when tested against a DTD. A valid XML document must be both well-formed and in conformity with a DTD. Tag set nesting, closed sets, case-sensitive language, attribute values, all the XML requirements that we had learned about at the beginning of the week, were now being brought into play in writing a tag set in conformity with the TEI DTD.

The TEI DTD is a predefined tag set used to describe the structural, administrative and descriptive elements of a document. Again Seaman emphasized project planning and management issues: value of element inclusion versus the cost of tagging; which elements of a document need to be tagged: names, dates, citations, variants, etc?

The importance of the TEI header as the location for bibliographic information was not startling to the librarians (shades of cataloguing course(s) at Library School), but the text elements of the DTD probably gave the class more trouble. Here we stopped being mere 'recorders' of information, and instead were participants in the creation of editorial content.

Should the names of residences be marked-up (Roberts issued many letters from, 'The Rectory' in Fredericton, but also lived in Toronto and New York during periods of his life)? Dates were obvious candidates for data searching, as were letter recipients, publishers and publications. Should we interfere with the content by noting spelling variations, provide footnotes about the people, places, events or items mentioned in the text of the letter? In short, we produced 'value-added' content to the document.

The set-up for the course, each participant at a networked station in a computer lab, lends itself to self-learning combined with team-teaching. In

undertaking the mark-up tasks, we could call on the services of those with experience in literary genres, bibliographic work, students with background in historical and biographical research, and many with varied computer and electronic text skills.

At times the noise of 'ahhs', laughter and forehead smacking gave the air of a frivolous activity, but it was really people recalling earlier lessons, realizing element entry mistakes ('close that tag set'), and trying to make the lessons come alive on the page. This is where the classroom experience proved itself more valuable (and more popular) than on-line quiz instruction. The benefit of a hands-on course with an instructor present in the room became apparent when it came time to validate the marked-up documents against the 'Validator' (isn't that a hockey player?). Few of our documents escaped a single 'ERROR' message (the first that did won a hearty round of applause), and so again we learned from one another.

The setting for the course was the idyllic UNB campus. The Harriet Irving Library, housed in a colonial red brick situated on a hill overlooking Fredericton, welcomes visitors to a coffee bar on entering. The Electronic Text Centre was still in the construction phase of its new digs, but already taking shape as an airy, colourful set of rooms off a main conference area: one could also feel the buzz of activity that surely engulfs the space today. The Centre's Director, Alan Burk threw open his home and backyard for a pleasant barbeque during the week offering participants an opportunity to gain insights into electronic text work from Text Centre personnel in a social setting.

The Institute is a very rewarding experience. It coupled hard work with hands-on experience: I know that every participant came away from the course with a fuller understanding of the complexities and the great rewards promised by electronic text creation and distribution. I would like to thank CIDL for awarding me the Summer Bursary that gave me the opportunity of attending this Institute. ☉