Mako Hill

Stephen Harris

Origins of Reading

15 May 2002

Final Paper - Thoughts on Computers, Readers and Text as Data

Advances in computer and information technology during the twentieth century have given people the ability to accumulate, process, manage and interpret data in ways and on scales that were previously unimaginable. Engineers are already able produce and use machines that can process and manipulate *terabytes* of data in seconds.[1]Every credit card transaction, every item purchased at a store, every telephone call, every email, and every website visit is recorded, processed and filed away. New fields of engineering involving data storage, recovery, collocation, and management are acting as the front line in twenty-first century's technological revolution. We're a long way away from file cabinets and card catalogs. It's sobering to think that much, even *most* of this data, will never be seen by a human being. This data is collected by computers and processed by scripts and applications into statistics and reports before any human sees it. Supermarkets don't distribute loyalty cards because anyone at *Stop and Shop* wants to read *your* grocery list; they collect and use this information for demographics and directed marketing. Most data is stored to provide context so computers can make make educated and more appropriate responses in the future.

In traditional forms of human literature, writers manipulate words in an attempt to use shared symbolic and associative connections to communicate meaning. The author's job is to encode ideas–or data–in textual form for transmission, dissemination, and preservation. In turn, readers parse these words, process and connect them in the context of their own

[1]Several companies now make routers, the machines that direct Internet traffic, that can more than a terabyte–one thousands of gigabytes or 10^{12} bytes of Internet traffic per second.

associative and symbolic experience and attempt to reassemble, decode and interpret the resulting data and, ultimately, to pull associative or symbolic meaning from the text. For years computers have, employing the same types of symbolic and associative processes, read and written data in genres all their own. As technology progresses, machines are beginning to take a more active role in the reading and writing of more traditional literary forms as well.

However, even today, computers' fundamental role in literature is largely that of glorified typewriter although intriguing advances have begun to challenge this paradigm. Powered by Microsoft Word, WYSIWYG word processors, and systems of *procedural* markup, computers act merely as cheaper, more effective, and more compact letterpress print shops; they treat literature as little more than marks on a page. Using software like TEX, LATEX, LYX, DocBook SGML, and other systems of *descriptive* markup, humans are beginning to use computers to approach traditional forms of literature as data–often similar to the way that humans do.[2]Acting as digital equivalents of designers, preprocessors, editors, and interpreters, these tools read and write literature in ways that historically have been distinctly human.

A text's materiality is both the result of interpretive spin and a major factor in the way the text is read and interpreted. The Hebrew Torah and the Christian New Testament are different texts formed from the same words–read and interpreted in respective scroll and codex forms, the differences in resulting religious philosophy and tradition are staggering. However, new technology is challenging old conceptions of literary materiality. Classified as data, text in DocBook SGML is edited in source form and then simultaneously published in a number of digital and print formats ranging from plain text to HTML to meticulously

[2]More information on TEX, LATEX, LYX, WYSIWYG word processing is available in a short paper I've written on the subject available online at: http://yukidoke.org/~mako/writing/origins/OR-Lyx.pdf More information on DocBook SGML and procedural versus descriptive markup is available in a longer paper at: http://yukidoke.org/~mako/writing/origins/OR-Markup.pdf This essay assumes familiarity with each of these concepts.

formatted printed pages. *O'Reilly and Associates* and other other publishers of technical books already insist that their authors employ descriptive markup systems to facilitate simultaneous publishing in digital and print forms.

Existing as codexes, digitized scrolls, and source data simultaneously, the materiality of a DocBook document is unclear. Is DocBook *every possible* material form? Is it every way of rendering the text: past, present and future? While material medium has historically played a major role in the way that texts are read and engaged with, digitized descriptive markup makes these super-textual elements dynamic in unprecedented ways. Technology has destabilized and threatened traditional approaches to material context in literature but has not depreciated its importance or effect. What this means is still unclear.

DocBook is designed to be explicit, clear, and unambiguous. It must be human writable and machine readable. While unprocessed DocBook source exists as a text on it's own, this text is not intended for human consumption. In most systems of descriptive markup, humans write texts for computers. Computers–after reading and interpreting these texts in the context of explicit rendering instructions–rewrite the literature for human readers. In this process, rendering software will discard data irrelevant to a particular material form or style–hypertext links make little sense on printed documents and emphasis won't show up in plain text. In determining how these texts are presented, rendering software must make important decisions about how the text will be presented and read–footnotes or endnotes?; readers will interact and understand each differently. To make these decisions correctly, the software must be able to read and *understand* the source text. By doing so, the computer joins, and often supersedes, the source's writer in the roles of editor, designer, and author.

Intriguingly, this relationship strains terms like editor, author and reader as the interaction between the source's author, rendering software, and the human consumer is unlike those that exist in non-digital contexts. DocBook is an example of how, critically embraced, computers can complement humans' abilities in interacting with literary texts. Fulfilling

roles that are difficult or impossible for humans, software like LyX and DocBook are playing an important roles in the evolution of the literary process.

Currently, computers (or at least personal computers) have trouble parsing and *understanding* data in the implicit way that humans can. As a result, cumbersome and overly-explicit systems like DocBook are needed to clearly convey what most humans can parse in very nuanced ways. Computers work best with explicit <*start*> and <*/end*> tags to unambiguously *know* where a chunk of data starts and ends. While even young children can tell from indentation where a block quote starts and finishes, computers currently prefer the text enclosed in <*blockquote*> tags as a more explicit form of data classification. While rendering from a printed page into DocBook is not impossible, it is a much more complex and error prone process–but it's one that human readers can do relatively easily. Through established systems of DocBook rendering instructions, or stylesheets, computers already have the information they'll need to preform this type of reading. And, over time, machines can and will learn to read non-textual data in these more implicit ways.

Unsurprisingly, there are already some computers that *do* read like humans and they are becoming increasingly common and increasingly advanced. Government intelligence systems like Echelon *listen* to international phone conversations and and *read* Internet traffic searching for "dangerous" phrases and ideas to bring to the attention of security and intelligence agencies. While every civil servant in every government agency could not begin to scan the bulk of trans-oceanic data, a basement of computers in Fort Mead running appropriately complex software can easily solve the problem. As computers become faster and cheaper, their power to read will filter down to educational institutions and to the general public.

While images of Echelon and a world of intelligent supercomputers are easily associated with apocalyptic messages, it's important to emphasize the fact that my vision is not an apocalyptic one. Computers will not make reading irrelevant any more than they will make

human beings outdated. Digital literary technologies like the World Wide Web have already allowed the transmissions of information about Echelon in ways that most traditional media sources have strictly avoided. The technology has allowed groups to stategize and carry out actions against the system.[3] While no technology is without its pitfalls, new technology approaching literary texts as data has the potential to revolutionize human reading in beneficial ways. As computers become better readers, they will be able to read millions of texts in the context of each other and help us trace symbolic and associate pathways between wide varieties of these works. Additionally, they will ask us to reflect on, refine, and redefine our concepts of reading.

[3] *Jam Echelon Day* has been organized over several years to raise awareness and decrease the effectiveness of communication monitoring systems like Echelon. More information is available at: http://www.cipherwar.com/echelon/ Similar efforts have raised awareness about freely available tools like Pretty Good Privacy (PGP) and the GNU Privacy Guard (GPG) which can secure data from spy systems like Echelon and other prying eyes.