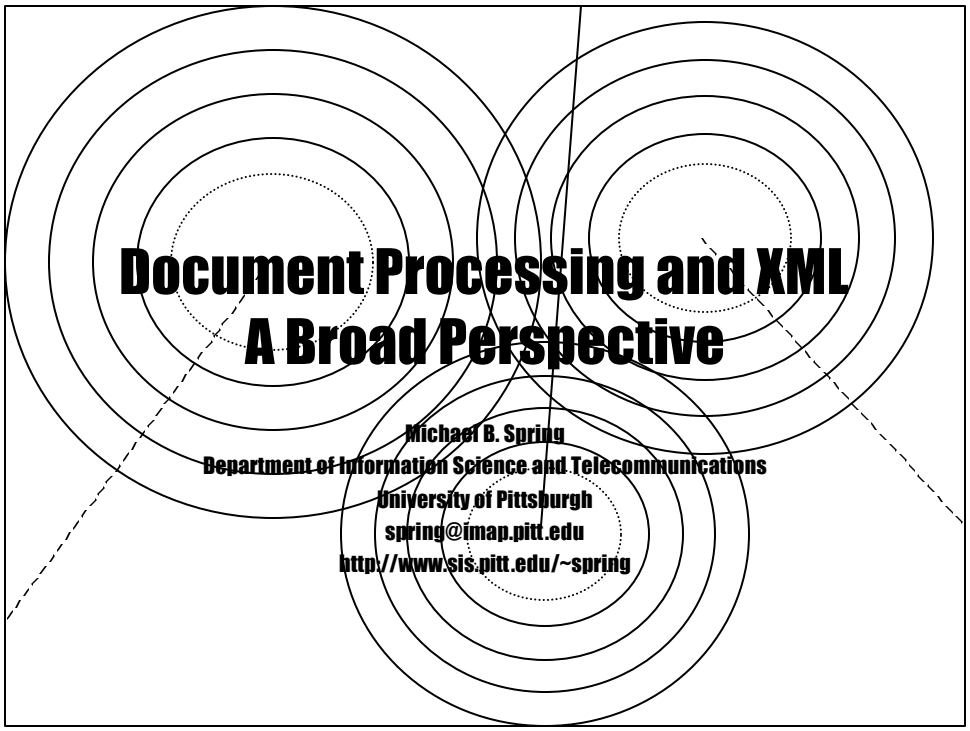


Prelude – Incunabula Revisited

- XML is the most recent form of efforts beginning with Pub and ru noff strengthened by GML, Scribe and XICS formalized in SGML, popularized in HTML.
- XML is the separation of logical and presentation structure with content situated in a directed acyclic graph.
- I don't know where the merger of documents and data that appears to be the destiny of XML will lead.
- I don't know where the morass of indeterminate style definition
  - Will style be creator defined
  - Will style be device defined
  - Will style be user defined
- I don't know how the merger of XML trees and Hypertext webs will play out.

Overview

- Perspectives
  - Personal history
  - Reflection points
- Overview
  - The history of reprographics
  - The computer and the document
  - Document processing matrix
- Where are we today
  - The Web
  - Stability
  - Capability
  - Dynamics
- What are the goals of the effort
- What role does XML play
- Next Steps

Personal History (Document Research)

- 1980: The Xerox STAR and academic publishing
- 1985: XICS, Planet Earth and custom publishing
- 1987: SGML and the Unstructured Text Converter
- 1991: Electronic Printing and Publishing: The Document Processing Revolution
- 1992: Hands on Postscript
- 1993: Mapping Abstract Data to Virtual Spaces
- 1994: CASCADE
- 1996: Balloting, Commenting, and Document Construction
- 1997: Multi-level Navigation of Document Spaces
- 1999: Social Awareness Tools

Reprographics Revolutions

- 1400-1600: Mass production (Y=cost/setup, X=cost/copy)
  - Block (a master to make copies)
  - Moveable type (a component based master)
- 1900-1960: Photo-optical processes (Y reduced twice)
  - Lithography (atomic level components, content neutral)
  - Xerography (reusable master)
- 1960-1990: Electronic processes (no Y, X distributed)
  - Fax (separation of master from copy)
  - Laser printers (elimination of physical master)
- 2000-????: Ad hoc reprographics (X eliminated)
  - WWW (elimination of physical copy)

Computers and Documents

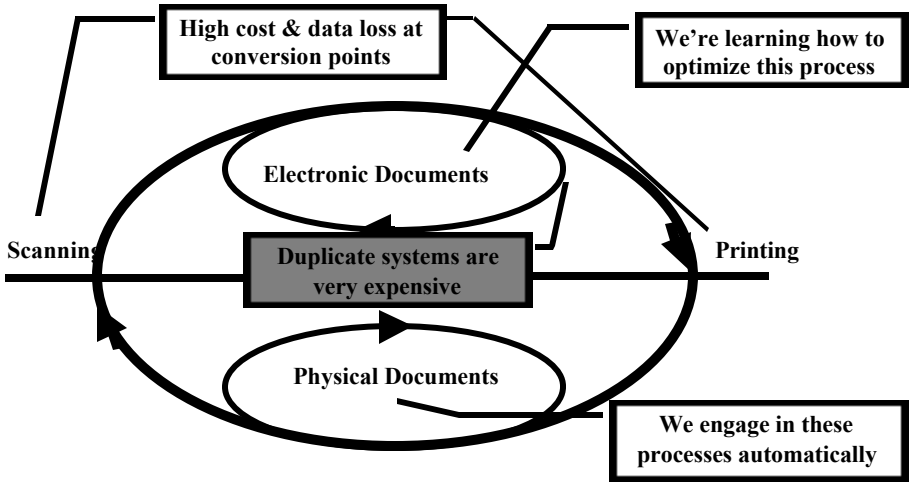
- Computer aided publishing or printing (1950-1990…)
  - Electro mechanical typesetting
  - Optical typesetting
  - High speed laser printing
  - Desktop publishing
- On-line databases (1960-1980)
  - Authoritative repositories
  - Full text systems
- CD-ROM publishing (1985-1995…)
  - Local area network services
  - Personal libraries
- WWW (1995-…)
  - Distributed publication

A Couple Points to Ponder

- Transition Costs: Documents are every business's second business – 6-10% of gross revenues. Transitional duplicate infrastructures consume profits
- Atoms to Bits: Documents are containers for ideas. Sometimes the containers are as important as the ideas -- the Constitution; your birth certificate; a love letter. We don't yet have a culture for container free ideas.
- Here Today– Gone Tomorrow: Documents used for decision making are increasingly ephemeral, to the extent that they may be irreproducible.
- Gone Forever: Archiving and provenience are both more sophisticated and more difficult in an electronic world (millennia media and millennia formats)

October 1, 2001 XML in Context 7

The Situation Conceptually



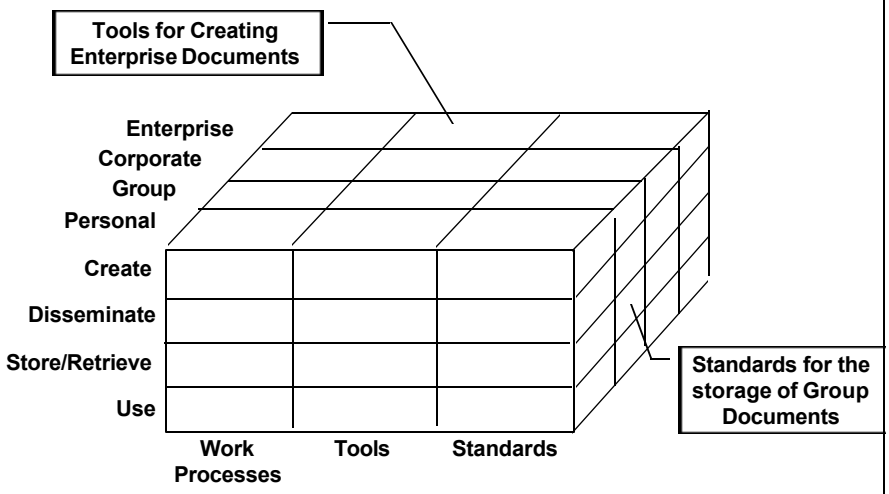
October 1, 2001 XML in Context 8

Important Document Processes

- Creation and Editing
  - text generation and format specification
  - Referencing, indexing, and illustrating
  - Interleaving and linking
- Storage and Retrieval
  - Classification
  - Association
- Distribution
  - Aggregators
  - Disseminators
- Use, Archiving and Disposition

October 1, 2001 XML in Context 9

Document Process matrix



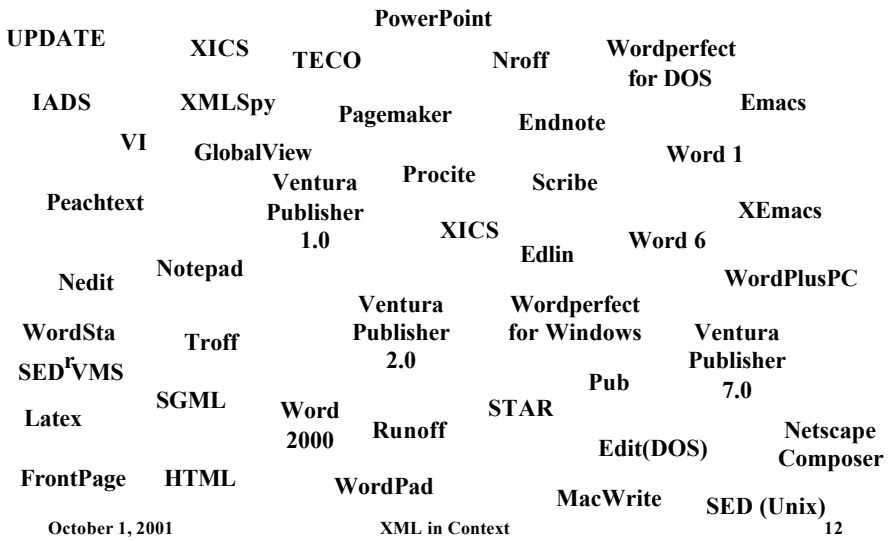
October 1, 2001 XML in Context 10

WWW and XML
"The End of the Beginning"

- The Internet provides a "stable infrastructure"
- Structured documents are accepted
  - Postscript and PDF
  - SGML, HTML, XML, and RDF
- Universal locators accepted
  - URLs, URIs, and URNs
  - PURLS and Object Object Identifiers
- New tools and document forms begin to emerge
  - Dynamic documents (scripted order forms)
  - Generated documents (catalogs and services)
  - Living Documents (reference materials and policy statements)
  - Personal Documents (ICAI and greeting cards)
  - Active Documents (voting queries, subscriptions)
  - Intelligent Documents (queries, advertisements)

October 1, 2001 XML in Context 11

Document Tool Stability



October 1, 2001 XML in Context 12

Document Tool Capability (provided)

Incredible Font Selections

Copy and Modify Styles

Metadata Attachments

WYSIWYG Editing

Inline Assistance and Correction

Multi-view Editing

Integrated Text, Graphics, Tables, and Images

Extensive Typographic Capability for Hardcopy Output

Global Search and Replace

Pattern matching

Extensive Conversions

Scripting Languages for Automatic Text

Version Control and Edit Tracking

Libraries of Boilerplate

Multiple Simultaneous Documents

October 1, 2001 XML in Context 13

Document Tool Capability (missing)

Standard copymarks across document types

Guaranteed Style layout that communicates a message

Quality Indexing

Automated sketching tools

Archiving of Documents to a Standard form

Bibliographic tools that know what to do

Standard document access information – e.g. headers and footers

Greatest Common Multiple Conversions – i.e. not Least Common Denominator

Within document locations (para 3 of page 22)

Automation of Multiple form documents – e.g. speech, slides, handouts, paper

October 1, 2001 XML in Context 14

Changes to be Accommodated

- Increasingly frequent revision
- Creation by copying and modification
- Distributed component documents
- Increased wide area collaboration
- Lack of presentational stability
- Distribution of the knowledge store
- Review and validation process eliminated
- Obfuscation of the copyright, intellectual property and ownership issues

Goals for Document Processing

- Refine input systems to move ideas to electronic form:
  - Making component building easy
  - Conversion of speech to formal exposition
  - Conversion of sketches to formal notation
- Establish a stabile electronic infrastructure for:
  - Storing and finding
  - Archiving and provenience
- Develop tools to
  - Index and filter
  - Register and archive
- Stabilize syntactic and semantic models for
  - Construction
  - Presentation
  - Query

The XML Model

- Structure, content, and presentation can be separated
- The structure of a document is a
  - A directed acyclic graph
  - Structural(logical) root branches to structure
  - Layout root branches to page sets, pages, and blocks
  - Content at the leaf nodes
- The header (DTD) provides a parseable/extensible definition
  - Prolog defines allowable instantiation and semantics
  - Prolog defines element attribute requirements
- The body (document instance) provides a highly structured set of labeled nodes
  - The nodes may be variously described

October 1, 2001 **XML in Context** **17**

One Agenda for Action

- Regain appropriate control of visual presentation as a part of the information transfer
- Make use of the attribute capabilities in XML to make the nodal components of documents richer
- Provide better tools to allow a casual user to make effective use of DTD's to instantiate rich, powerful, stabile, personal, and productive documents
- Develop tools that make use of visual skills to recognize structure and navigate document spaces ranging from individual documents to archival collections
- Work to create a social periphery in the document space that brings humans closer together

October 1, 2001 **XML in Context** **18**

Regaining Visual Information

- The 1980's were the Golden Age of visual information.
  - Pagemaker and Ventura provided everything from tracking to complex hyphenation to running headers.
  - The media presentation could enhance the substantive message at an incredible level of detail
  - Laser printers exceeded the 480dpi resolution
- In the 1990's ad hoc reprographics dramatically increased distribution reducing presentation quality
- Beautiful page design features have been lost
- A new approach to presentation settings is needed:
  - What is author, user, and device defined
  - Intelligent visual definition of presentation
  - Ad hoc display devices have to standardize

October 1, 2001 **XML in Context** **19**

Flesh Out the Nodes

- The Alexandrian and other libraries created a need for document level identification – e.g. title pages
- SGML and ODA offered great promise of providing attribute information would add much clarity to structured documents -- each node would have an idea, an author, and numerous other attributes specified
- Nodal attributes must be expanded
  - Information about the author, origin, and revision of nodes must be captured automatically
  - Possible uses need to be explored and standards developed that will encourage use
  - Systems for visualization of the data need to be worked out

October 1, 2001 **XML in Context** **20**

Creation of Document Instances

- Historically, authoring has been:
  - An ad hoc process
  - A linear process
  - An individual process
- Increasingly it is a structured group cyclic process
- New tools are needed
  - GUI instantiation of documents in accord with DTDs
  - Automated specification of attribute data by scripting
  - Protection of documents and document components via inherited access control lists
  - Branch pruning and grafting for collaborative authoring
  - Version control tools for selective reconstruction of documents

Navigation of Document Spaces

- Historically, we have relied on libraries and journals to help us navigate document spaces
- We need new tools to navigate associatively organized spaces
- Visual overviews of spaces
  - By structure
  - By attribute
  - By change
- Usage linking of objects
  - Collaborative filtering
  - Latent semantic indexing

A Sense of Place in Space

- A feel for document goodness
  - Am I done writing this document
- A feel for author involvement
  - How is the collaborative effort going
- A feel for document value
  - How is this document valued by others
    - Authoritative others
    - Peers
    - Whoever