The Document Processing Revolution

Michael B. Spring
Department of Information Science and Telecommunications
University of Pittsburgh

Overview

- Prelude
- History
  - People
  - Three perspectives
- Rationale
- Concepts and terminology
- Technologies and tools
- Models
- Futures
  - Changes
  - New Forms
- Conclusions

Prelude

- I’ve been studying documents for 20 years
  - What is a document?
  - Name a document process?
  - What new forms of documents are there
  - What new tools are needed
- My head hurts
  - XICS/Scribe/nroff/Ventura/Latex/Word
  - XML/XSL/XSLT/XLL/XPath/XQL
  - xt/DOM/SAX/jaxp/xlan/

Documents -- Conceptual

A document is an identifiable entity having some durable form, produced by a person or persons toward the goal of communication; it may take a number of forms, but must have a least one symbolic manifestation that used to store or communicate information between people. It is a cohesive entity formed of subcomponents in logical, layout, and content form.

Documents -- Descriptive

- There was a day a document was a report or a book that consisted predominantly of text written by a single author.
- Documents today are no longer so simple, they:
  - Include text, graphic, images.
  - May be cohesive, e.g. a letter or a report
  - May be a loose composite, e.g. a medical record.
  - May be authored by individuals, groups, or organizations.
  - May have a limited life span or be archival.

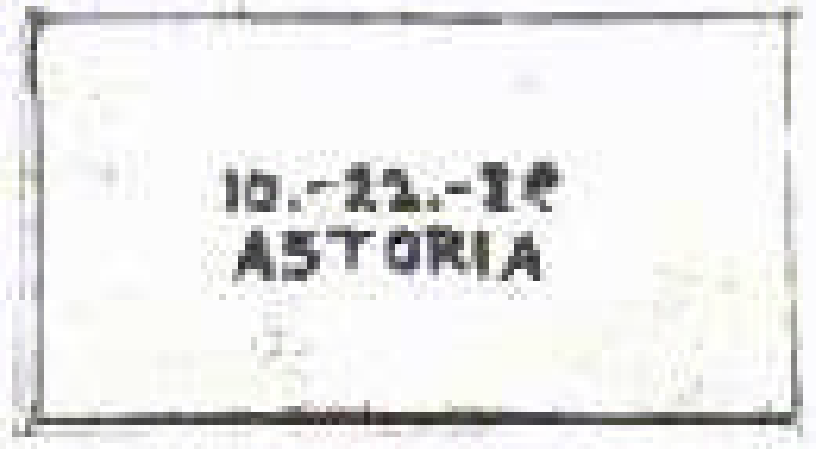
Six Stories

- Six stories told
  - Ptolemaios Soter
  - Chester Carlson
  - Vannevar Bush
  - Douglas Engelbart
  - Alan Kay
  - Brian Reid

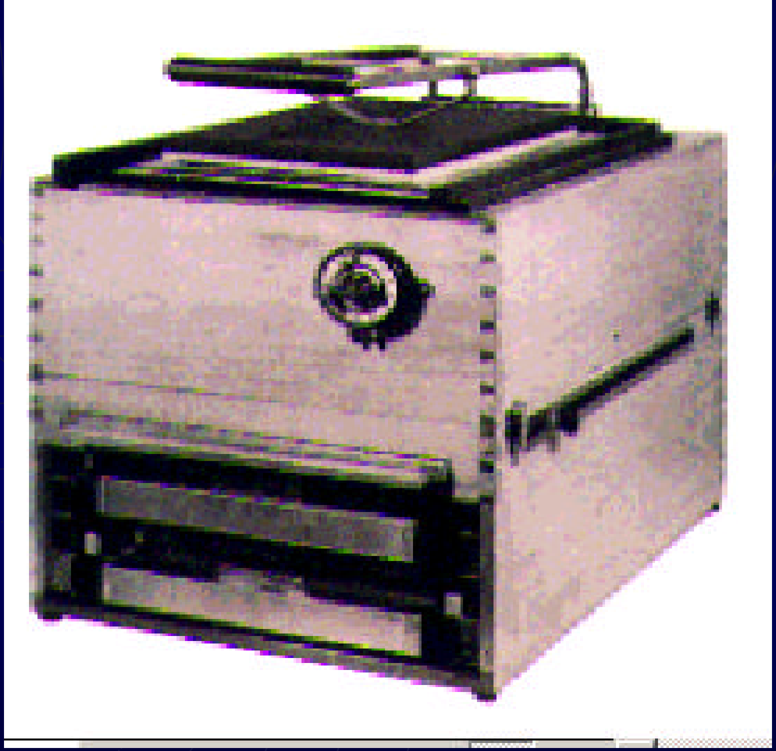
Ptolemaios Soter and Buckets

- Demetrios Phalereus persuaded Ptolemy I (Ptolemaios Soter) in 307 B.C. to collect copies of all known books to be placed in Alexandria
- This repository flourished for many centuries eventually amassing over 750,000 scrolls and papers on a wide range of subjects.
- The first organization of topics was probably modeled after Aristotle and included:
  - mathematics
  - medicine
  - astronomy
  - Geometry

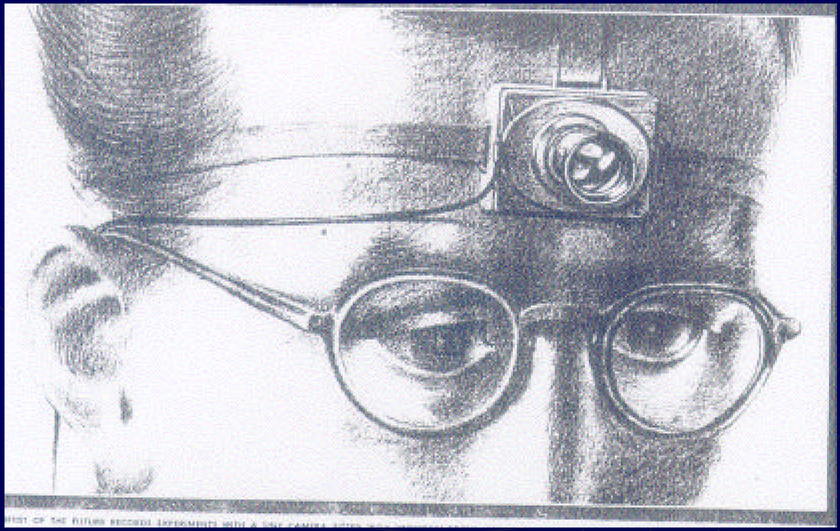
Chester Carlson and Xerography

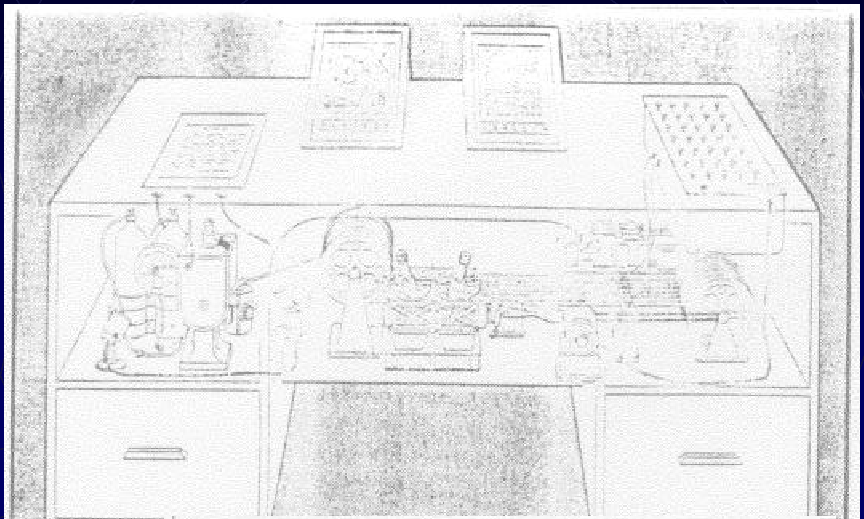


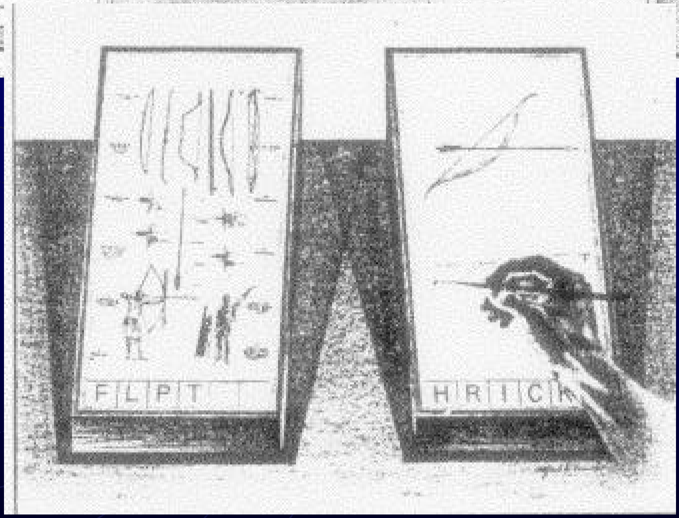
10.-22.-38
ASTORIA



Vannevar Bush and Association







Douglas Engelbart and Augmentation

Alan Kay and Dynabook

Brian Reid and Scribe

```
\cpi12,propon,lm5,lw80,tm6
\bm6,bf3,cnp3,pi6,sp1,justc
\ctr\@Faculty Development Presentation
\ctr\January 26, 1984
~Introduction:I will cover three topics:
     First, the reasons why we should be thinking about tv
    Second, some of the things to keep in mind in working
with video
    Third, some ways to get started
~Reasons:  We should be looking at video
because:
     The influence of Walter Annenberg and Mobil Oil
    The emergence of TAGER and PECS
    The growth of cable -- implications of over channeling
    The increase in satellites -- implications of abundance
    Microcomputer controlled videodiscs -- a marraige made
in heaven
\np
~How to get started
```

```
@make(report)
@begin(titlepage)
@title[COMPUTER CENTER REPORT]
@date[January 12, 1984]
@end(titlepage)
@chapter(DEPARTMENTAL LIBRARIES)
The library for Computer Science, CSL:, has been created,
with a quota of 10,000 blocks.
Free space on SPL: was critical during the Fall term. It is
currently at 106,000 for System A and 122,000 for System B,
and will decrease rapidly as the Winter term progresses.
@section(INFORMAL COURSES)
The schedule of informal courses for the Winter term has
been announced.The courses being offered are
@begin(list)
Computing for the New User
Introduction to Graphics at Pitt
Interactive System 1022
@end(list)
Please see SYS:NEWS for details.
```

Digital Document Processing “Revolutions”

- 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

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)

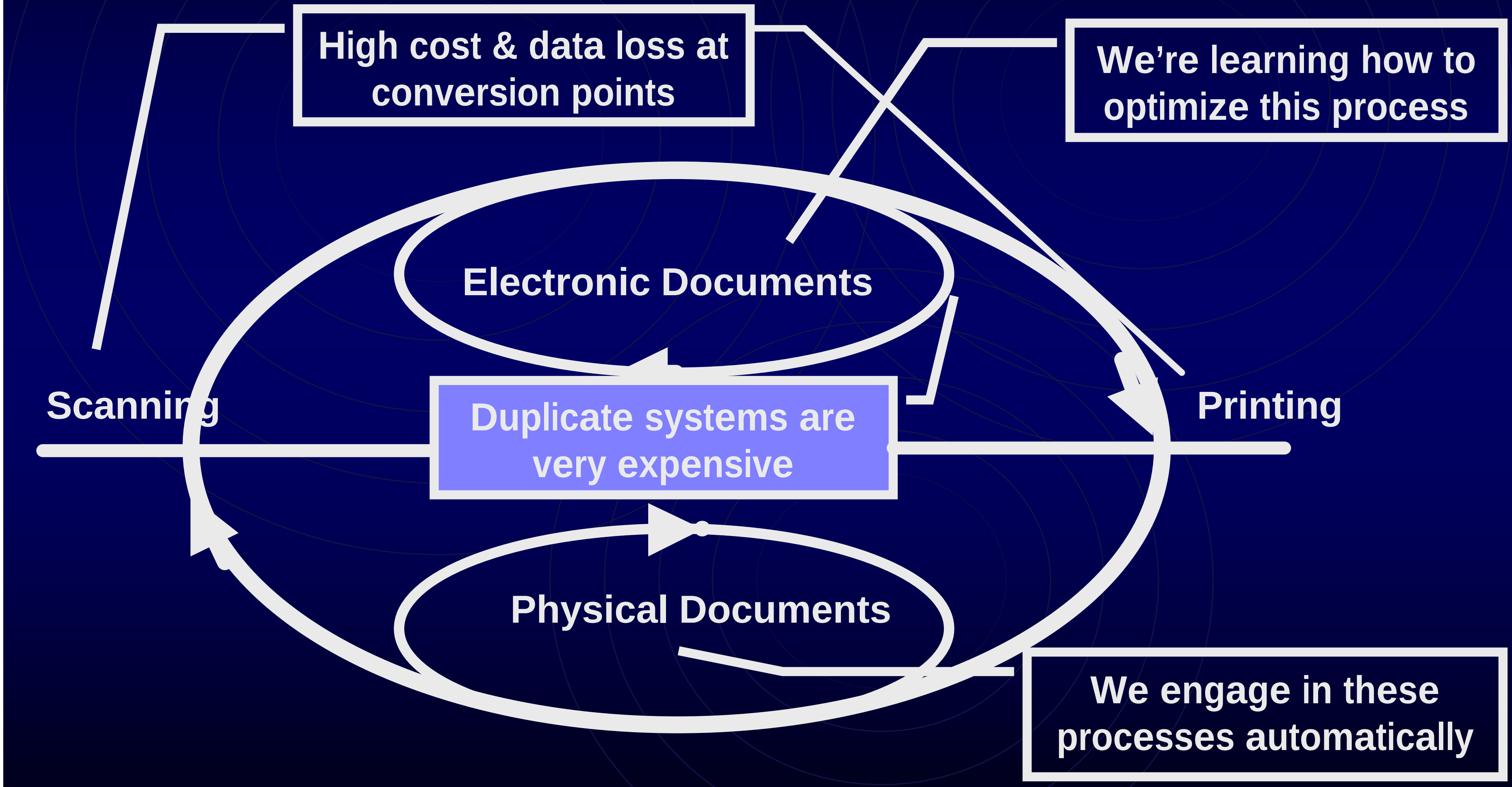
Communications Revolutions

- The oral tradition (50,000)
  - Knowledge dies with the bearer
- The written (literary) tradition (5000)
  - Knowledge across time and space
- Telecommunications tradition (150)
  - The second orality
  - A non intermediated instantaneous communication
- Computer mediated tradition (5-50)
  - The second literacy
  - All prior forms plus active intelligence

A Couple Points to Ponder

- Transition Costs
  - Documents consume 6-10% of gross revenues.
  - Transitional duplicate infrastructures consume profits
- Atoms to Bits
  - Documents are containers for ideas
  - We don’t yet have a culture for container free ideas.
- Here Today– Gone Tomorrow
  - Documents used for decision making are increasingly ephemeral
- Gone Forever
  - Archiving and provenience are both more sophisticated and more difficult in an electronic world (millennia media and millennia formats)

Duplicate Infrastructures



High cost & data loss at conversion points

We're learning how to optimize this process

Electronic Documents

Scanning

Duplicate systems are very expensive

Printing

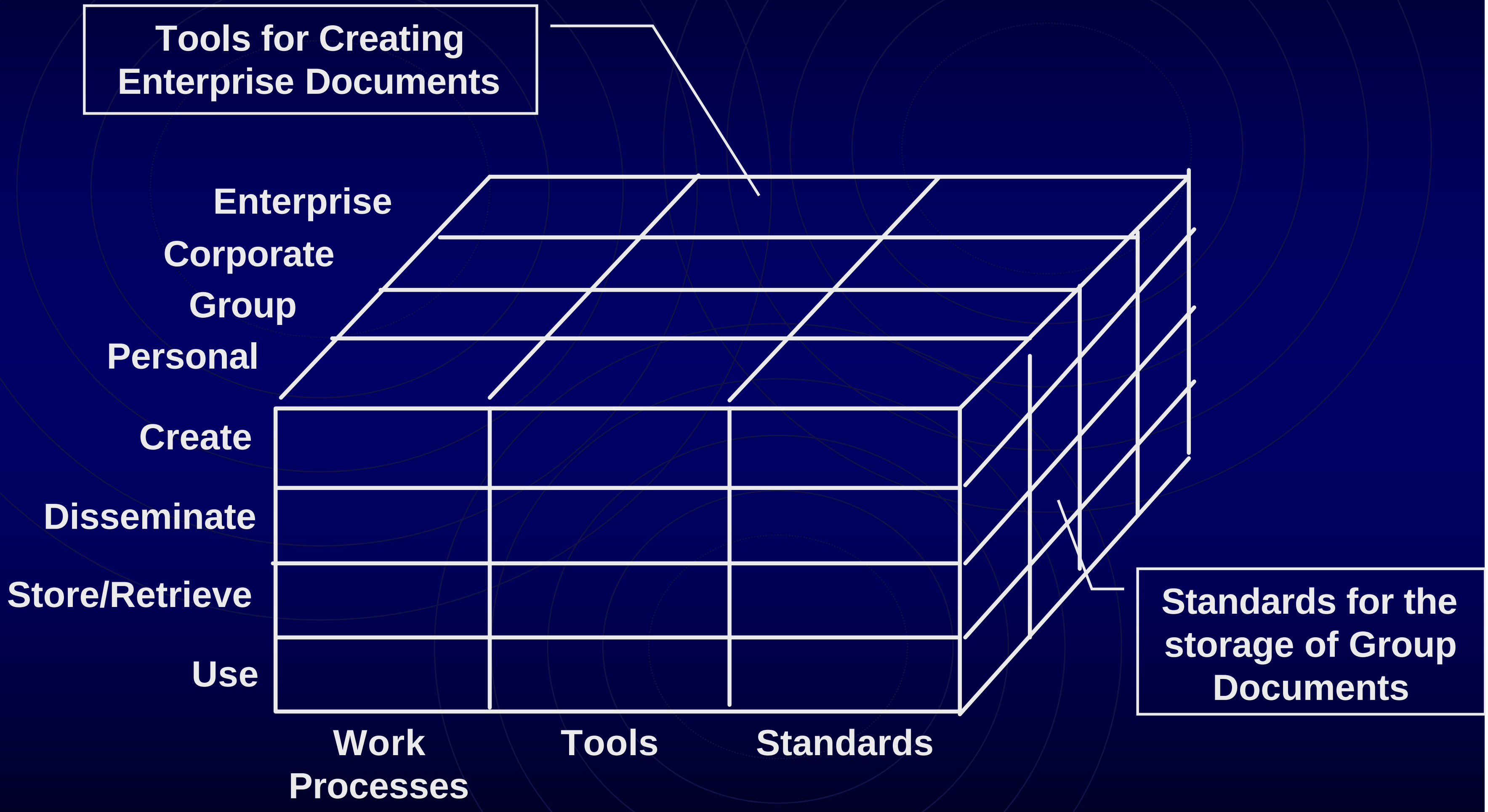
Physical Documents

We engage in these processes automatically

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

Document Process matrix



Tools for Creating Enterprise Documents

Enterprise

Corporate

Group

Personal

Create

Disseminate

Store/Retrieve

Use

Work Processes

Tools

Standards

Standards for the storage of Group Documents

How is the Transition Going

- Creation and Editing
  - Creation tools favor text over graphics, equations, etc.
  - Modeless editors have matured
  - Universal functions exit
- Composition
  - Structural composition and editing is weak
  - Conversion and transformation is maturing
- Rendering
  - Rendering presents new dilemmas related to control
  - Classic presentation problems are now reconsidered
    - Hyphenation, running heads, widows and orphans
    - Long footnotes and juxtaposition of text and graphics (exterminate versus exotic foods)

WWW and XML
“The End of the Beginning”

- The Internet provides a “stable infrastructure”
  - IP/TCP
  - DNS
- Structured documents are accepted
  - SGML, HTML, and XML
- Universal locators accepted
  - URLs >> URIs
  - PURLS and Object Identifiers
- Meta languages are emerging
  - Resource Description Framework
  - Topic Maps
- Directory Services are emerging
  - URNs
  - LDAP

Goals in Document Processing

- Understand the process
- Develop the conceptual infrastructure
  - Semantic and architectural models
  - Structural and syntactic processing
- Develop “as good as” digital tools
  - Authoring and structuring
  - Indexing and filtering
  - Rendering
- Foster cultural acceptance of new forms
  - Ownership (copyright)
  - Migration to structured format

Understanding the Old Process

- Measures
  - points, ems and x-height
- Appearance
  - loose, serif, oblique, normal
  - Gutters, columns, and page position
- Images
  - Dithering, screens, and anamorphic scaling
- Components
  - Headings, indices, cross references, and footnotes
- Processes
  - Publishing, typesetting, and printer

Using Technology: Scanning as an Example

- We are concerned with scanning for a minimum of three reasons
  - Preparing images
    - What is equivalent
    - How is it produced
  - Recognizing text
    - Recognition accuracy
    - Recognition scope
  - Preserving analog documents
    - What is needed when
    - How is provenance assured

Image Reproduction

- Photographic screen resolutions for half tones allows infinite variations on each dot
  - A 200-300 line screen for fine art requires a digital resolution of 800-1200 dpi
- Digital processes simulates screens by dithering
  - Dithering formulas – dot centered, dot dispersed, error diffusion – produce different results
- Image Transformations include
  - point functions -- bais (lighten/darken), gain (contrast), histogram equalization
  - area functions -- smoothing, noise reduction, etc.
  - filters

Recognition Techniques

- Character recognition based on
  - pattern matching
  - feature detection
- Segmentation and skew correction to recognize
  - lines
  - columns
  - images
  - line drawings
- Document recognition which uses all of the above to identify semantic components – title, author, etc.

The Evolutions of Document Processing Models

- Computer science model
  - Input >> process >> output
  - Byte/bit level operation
  - Line and column view
- Office Automation Model
  - Tasks and users
  - Strings
  - Sentence and paragraph view
- Document Processing Model
  - Structural model
    - Trees (SGML/XML)
    - Network (Hypertext)
  - Objects
  - Nested Object view

“New” Concepts

- Tree Structures
  - Logical tree – document content models
  - Layout tree
- Hypertext
  - Anchors
  - Links
  - Nodes
- Document Content Modeling (SGML/XML)
  - Elements
  - Attributes
  - Entities

A Basic Model for Electronic Documents

- Tree
  - Logical root branches to structure
  - Layout root branches to page sets
  - Content at the leaf nodes
- Metalanguage for description
  - Prolog defines instantiation structure
  - Text and markup can be validated
- Standards for
  - Markup
  - Content modeling
  - Navigation
  - Transformation

Selected Changes

- Cultural changes
  - Frequency of revision
  - Interactive composition
  - Visual text -- format communicates
  - Hacker attitude toward ownership
- Structural Changes
  - The process flop
    - Disseminate then print
    - Publish then review
  - The structure explosion
    - Scripting
    - Integrated graphics and data
  - The storage and retrieval paradigm shift
    - Keywords to full text to topicmaps

Future Document Forms

- New forms will meet special needs
  - Rapidly Changing Documents (reference manuals)
  - Dynamic documents (scripted order forms)
  - Generated documents (catalogs and services)
  - Living Documents (reference materials)
  - Complex Documents, (standard sets, encyclopedias)
  - Ephemeral Documents, (policy statements)
  - Multimedia Documents, (journals, manuals)
  - Personal Documents, (greeting cards, training manuals)
  - Active Documents, (voting queries, subscriptions)
  - Intelligent Documents, (queries, advertisements)

Conclusion (1991)
There is going to be a change

We are blind if we think that the publishing house or print shop, both prime examples of the institutions of a passing industrial society, will not undergo a radical transformation. The immediate economic effects, already taking place, will be felt by the specialized work force which will become obsolete as their jobs are redefined. The shape of the institutions themselves must change as well, especially as we begin to see the functions of printer and publisher move into the corporation, the university, the library, and the home. As these other segments of society take on new roles in the processing of information, these institutions will also change. (p 54 of EPP/DPR, 1991)

Conclusion (2001)
Powerful forces are afoot

- Economics
  - Cost reduction through reusability
  - Costs born by the user
  - Use versus ownership models
- Technology
  - Eye level displays – printing and screen
  - Ubiquitous connectivity
  - An accepted standard set
- Culture
  - Structured document forms
  - Collaborative posture
  - New forms of storytelling