Introduction to Document Management

Documation ‘97
February 25, 1997
Santa Clara, California

Kurt Conrad
conrad@sagebrushgroup.com

Goal of Tutorial

To help you to understand the fundamental changes which are occurring in the field of document management and their relationships to process and technology alternatives.

Fundamental Changes

- Just now learning to use computers to improve organizational performance.
- Destabilizing the nature of work
  - Organizational purpose
  - How individuals contribute value
- Document management “in the cross-hairs”
  - Concept of the document
  - Measures of value

Hidden Importance

- P 80-90% of corporate information in documents
- P Documents claim
  - < 40-60% of office worker’s time
  - < 20-45% of labor costs
  - < 12-15% of corporate revenues
- P Emerging metaphor for organizing complex information

Documents as Strategic Assets

P Contain information critical to complex organizational behaviors

< Provide context

< Integrate, document, and communicate understanding

P Critical to customer satisfaction

P Inconsistently recognized as strategic

< Real men do databases

< CALS, ATA 2000, ISO 9000, etc.

What the Tutorial Will Cover

P What is Document Management

P The History of Document Management

P Document Management Architectures

P Implementation Issues

P Workflow Automation

P Integration Points

P Impact of the World Wide Web

What is Document Management?

Simple Definition

Systems for managing collections of documents

Wide disparity of approaches

- Document Image Management
- Full Text Retrieval
- Compound Document Management
- Online Viewing
- Workflow
- Object-Oriented Databases

What is Management?

Actions taken today to protect the future

Protecting the Future

P Do all your documents (or the information in them) have the same future?

< “One size fits all” solutions are a common mistake

P How much will the future cost?

< Cost = (Legacy, Vision)

P Future value is defined in terms of human and automated behaviors

Metadata Determines Future Value

P Metadata = data about data

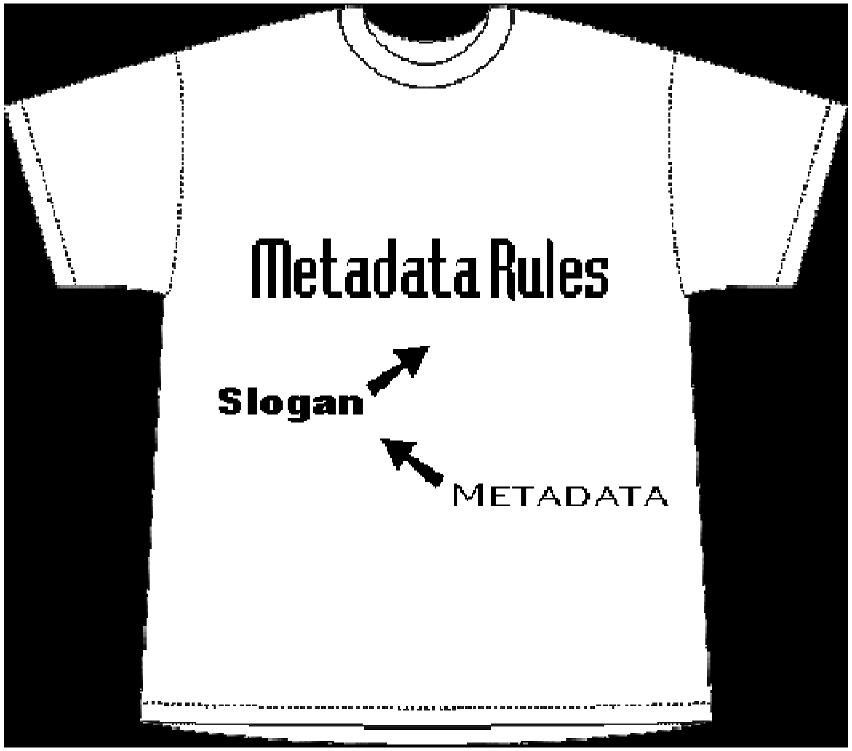
P Metadata is the basis for behavior

P Humans can create metadata and resolve ambiguous metadata

P Computers can’t

P Documents are often rich in ambiguous metadata

P Are your documents “smart enough” to meet future needs?



What is Document Management?

P Document Management processes and technologies protect the future value of documents.

P A wide variety of approaches have been developed which are based on different concepts of the document and emphasize different definitions of document value.

History of Document Management Systems

History Overview

P Mirrors the evolution of the concept of the document

P Conceptual changes closely tied to technology and metadata changes (chicken and egg)

P Three primary concepts

< Paper documents

< Automated paper documents

< Electronic documents

Paper Documents

Focus on the dynamics of the physical artifact

P Metadata implied through visual clues

- < Linear sequence
- < Typography and formatting
- < TOC, lists, indexes, cross references, etc.

P Human interpretation creates meaning

P Efficient use of space often more important than retrievability and reuse

P Innovations target the independent efficiency of production, storage, and retrieval

Automated Paper Documents

Speeds the processing of physical documents

P Paper hides a multitude of sins

P Focus on visual formatting

< Laser printers allow more control

< HW/SW tools function like fast, powerful pens

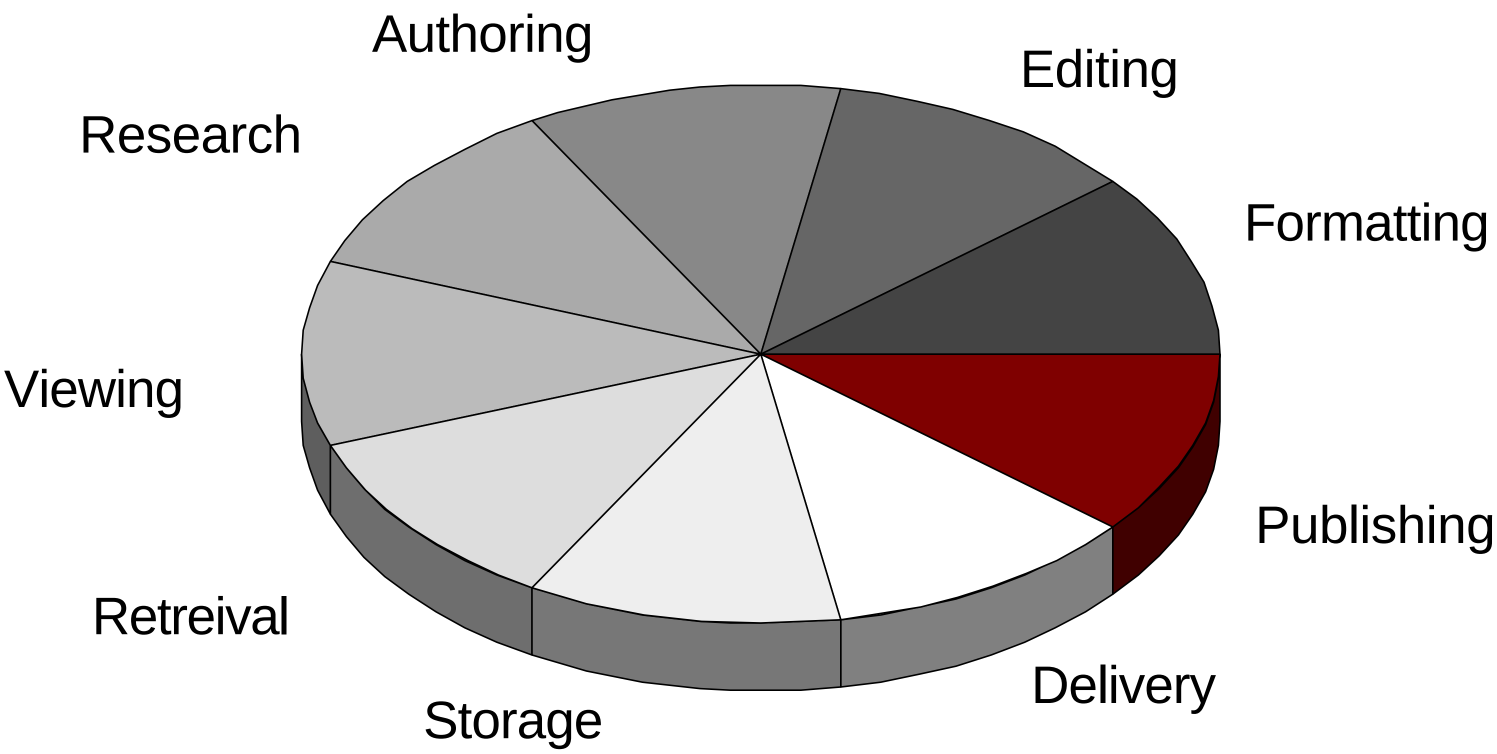
< Metadata / operator interaction based on formatting codes

P Illusion of control

P Management of meaning and semantics limited to relational database world

Automated Paper Documents

Solutions often focus on a subset of the document lifecycle



Automated Paper Documents

Technologies

P Paper-based interface standards

P Graphics, Wordprocessing, and Desktop Publishing tools

P Manage information *about* the documents

< File management systems

< Image management systems

< Other database-based indexing systems

Electronic Documents

Conceptual Shifts

P Increased information density

P Documents are more than their paper representations

  < Time-based media

  < Hyperlinks and other navigational aides

  < Formal relationships to other sets of information

P Paper becomes a portable, high-resolution display technology

Electronic Documents

Conceptual Shifts

P Processing-neutral encodings that support multiple representations for delivery

P Emphasis on meaning and semantics

  < Richer, more descriptive metadata that serves as a basis for integrating the entire document lifecycle

P Tied to new organizational models that are based on shared pools of information

Electronic Documents

Performance

P Time and quality become dominate values

< Use and reuse of knowledge

< Customer satisfaction

P Performance and value increasingly limited by production process

P Increased importance of up-front design

< Formalized structures and validation

< Explicit metadata that supports complex human and automated behaviors

< Software and data interfaces

Electronic Documents

Technologies

- P Manage information *contained in* documents
- P Data encodings as interface standards
- P Structured authoring
- P Hypermedia authoring (including links, annotations, workflow, other relationships)
- P Component management systems
- P Convergence of competing concepts

What is Document Management?

Revisited

P Today’s high-performance documents are based on relationships

P Emphasis is shifting away from

< Simple storage and retrieval

< Independent management of life cycle phases

P New emphasis on integrating interrelated information lifecycles

P Systems often encompass competing concepts of the document

Overview of Document Management Architectures

Overview

- Three models
  - Image-based
  - WYSIWYG DTP
  - Compound document management
- Components
  - Data encoding standards
  - Software interoperability standards
  - Task-specific tools
  - Communications and repository infrastructure

Image-based Architectures

- Dragging paper documents into the electronic age
- Heavy reliance on human interpretation
- Layering of metadata to capture meaning and understanding
- Workflow automation and annotation innovations

WYSIWYG DTP

P Control of visual aspects

P File-based and BLOBS

P Production focus

P Short-lived documents

< Advertising
< Novelty
< Drama

P WWW

Compound Document Management

- Control of individual information objects
- Structure and semantics
- Late binding of typography
- Customization of both form and content
- Addressing and transformation issues
- Encompasses and consolidates other architectures

Data Encoding Standards

General Questions

- Who controls the standard?
- What classes of metadata (conceptual models) does it support?
- What behaviors does it support?
- Portability, platform independence, ability to support required transforms

Data Encoding Standards

Text

- Paper
- Image
- Text
- Page image
- Traditional markup
- Generalized markup

Data Encoding Standards

Graphics

- Paper
- Image
- Vector
- Semantically-rich vector graphics

Data Encoding Standards

Other

- P Audio
- P Video
- P Voice
- P Positional
- P Hyperlinking
- P Rendering
- P Behaviors

Software Interoperability Standards

- Programming languages
- Application Programming Interfaces
  - Single vendor
  - Vendor consortium
- Examples
  - Shamrock, DEN, ODMA, OLE, OpenDoc, CORBA
- Stability

Task-Specific Tools

Authoring

P Traditional

- < Word processing and DTP
- < Graphics

P Structured authoring

- < SGML/HTML
- < Forms
- < Graphics

P Layering

- < Browsers

Task-Specific Tools

Editing

- Heavily reliant on human interpretation
- Syntax checkers and validators
  - Content (spelling, grammar)
  - Markup
- Batch vs real-time

Task-Specific Tools

Formatting & Publishing

P Converters

< Scanners

< OCR/vectorizers

< Programmable

P Composition tools

P Physical media and associated hardware

P Hypermedia authoring tools

P Print on demand

Task-Specific Tools

Delivery & Storage

P Dependent on published form

P Relational and object-oriented databases

< Square pegs

< Tables, hierarchies, and non-linear relationships

< Performance

< Data model designs

< Granularity

P Email, workflow, other network-based transport mechanisms

Task-Specific Tools

Retrieval

- P Database queries
- P Full text
  - < Boolean searches
  - < Weighted thesauruses
  - < Vector searches
  - < Context-sensitive searches
  - < Natural language
- P Image matching

Task-Based Tools

Viewing

- Text readers
- Native file viewers
- Raster viewers
- Page viewers
- Binary browsers
- Fixed markup language browsers
- Arbitrary DTD browsers

Infrastructure

- Repository and communications subsystems
- Scope
- Granularity
- Encodings
- Versioning and configuration control
- Target of most software interoperability standards

Implementation Issues

Human Issues

P Difficulty of adopting enabling technologies

- < Conceptualization
- < Learning
- < Foresight

P Perceptions

- < Technology problem
- < Uniqueness

P Who knows?

Organizational Issues

P Reengineering

< Complex behavior based on richer semantics

< Self-awareness

P Information politics

< Stakeholder interests

< Policy development & governance

< Allocation of decision making

P Competing interests of information owners and technology vendors

Technical Issues

P Adequate communications infrastructure

P Cross-platform integration

P Selecting standards

P Legacy systems and data

P Addressing and granularity

P Planning for obsolescence

P Labor costs

Workflow Automation

Issues

- Often confused with document management
  - Check-in and check-out
  - Component-level configuration control
- Convergence with document management
  - Routing and communication
- Ad hoc vs engineered workflows

Opportunities

P Basic reengineering model

< Shift from linear flow to shared pools

< “Linear” process flows still remain

P Documenting transformations provides additional context to information objects

< Facilitates understanding

< Simplifies reuse in new contexts

P Additional “publishing vectors”

Integration Points

Organizational Integration

- Information suppliers and consumers
- Metadata requirements
- Process, policy, politics
- Values

Data Integration

- P Encoding standards
- P Software interoperability standards
- P Transformations
- P Addressing
- P Synchronization

Impact of the World Wide Web

Primary Impact

First time that a large number of individuals and organizations have used non-proprietary, vendor-neutral encoding and communications standards to implement a truly heterogeneous computing environment.

Additional Impacts

- Encoding standards
- Software design
- Focus for consolidation

Encoding Standards

P HTML hides a multitude of sins

P A application of SGML

< Conformance issues

< Volatility

< Theology

P Easy to get into

P Danger in thinking that more than a delivery encoding

Encoding Standards

P Simplicity limits utility and drives divergent publishing models

< Complex graphics

< Structured data at the server

P Competing/complementary efforts

< Stupid HTML export

< Proprietary encodings

< Increased visual sophistication

< Structural flexibility

P XML Initiative

Software Design

- Viewer-centric
  - Customized views
  - “Do everything” browsers
- Smaller apps (e.g., plug-ins, java applets)
- Platform independence
- Authoring metaphors

Focus for Consolidation

P Aim for the accident

P Change changes change

< Perceptions of value

< User needs

< Vendor desires

Conclusion

- Use encodings as primary integration mechanism
- Choose tools that let you control metadata structures and object granularity
- Layer new relationships and meanings as identified
- Engage stakeholders in all phases of document lifecycle to identify metadata requirements