Introduction to Document Management

Documation '98 West
March 9, 1998
San Jose, California

Kurt Conrad
kurt.conrad@documentum.com
conrad@SagebrushGroup.com

Documentum, Inc. ❑ 5671 Gibraltar Drive ❑ Pleasanton, CA 94588-8547
Phone 510-463-6800 ❑ Fax 510-463-6850 ❑ http://www.documentum.com

Where I Come From

- P Senior Consultant with Documentum, Inc.
- P Founder: The Sagebrush Group
  - Independent consulting (95-97)
  - Professional association of practitioners
  - http://www.SagebrushGroup.com
- P 14 years Boeing Computer Services and the Department of Energy

Documentum, Inc.

Corporate Focus

Develops, markets, and supports a family of client/server and web software products that enable companies to share, manage, and reuse the vital corporate knowledge contained in business-critical documents.

Documentum, Inc.

Enterprise Document Management System

P Used by >350 of Global 1000 companies

P Designed for

- ‣ Rapid and flexible deployment
- ‣ Ease of use
- ‣ High return on investment
  - – Accelerating time to market
  - – Improving product quality
  - – Enhancing operational efficiency
  - – Insuring compliance

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.

Is this tutorial about SGML/XML?

P No

- ▸ The Standard Generalized and Extensible Markup Languages are explicitly mentioned only a couple of times
- ▸ Key issues have little to do with technical aspects of SGML and XML

P Yes

- ▸ Been involved with the SGML since 1992
- ▸ Colors all of my thinking about documents
- ▸ Logical conclusion to emerging strategies of reuse
- ▸ XML very likely to be central to next-generation tools

Fundamental Changes

P Just now learning to use computers to improve organizational performance.

P Destabilizing the nature of work

- Organizational purpose
- How individuals contribute value

P Document management “in the cross-hairs”

- Concept of the document
- Measures of value
- Revolutionary technologies

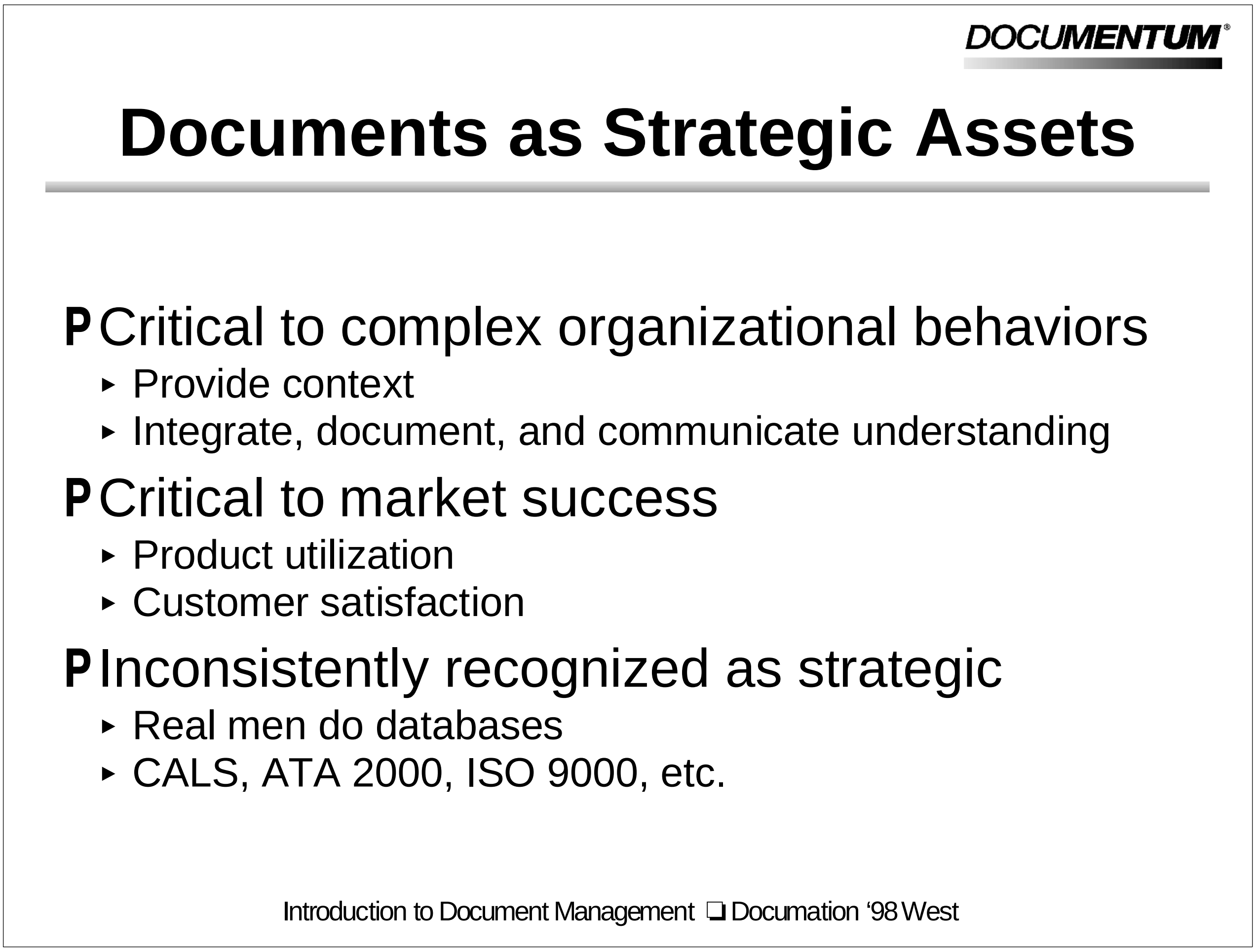
Hidden Importance

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 Docments have become the emerging metaphor for organizing complex information



DOCUMENTUM®

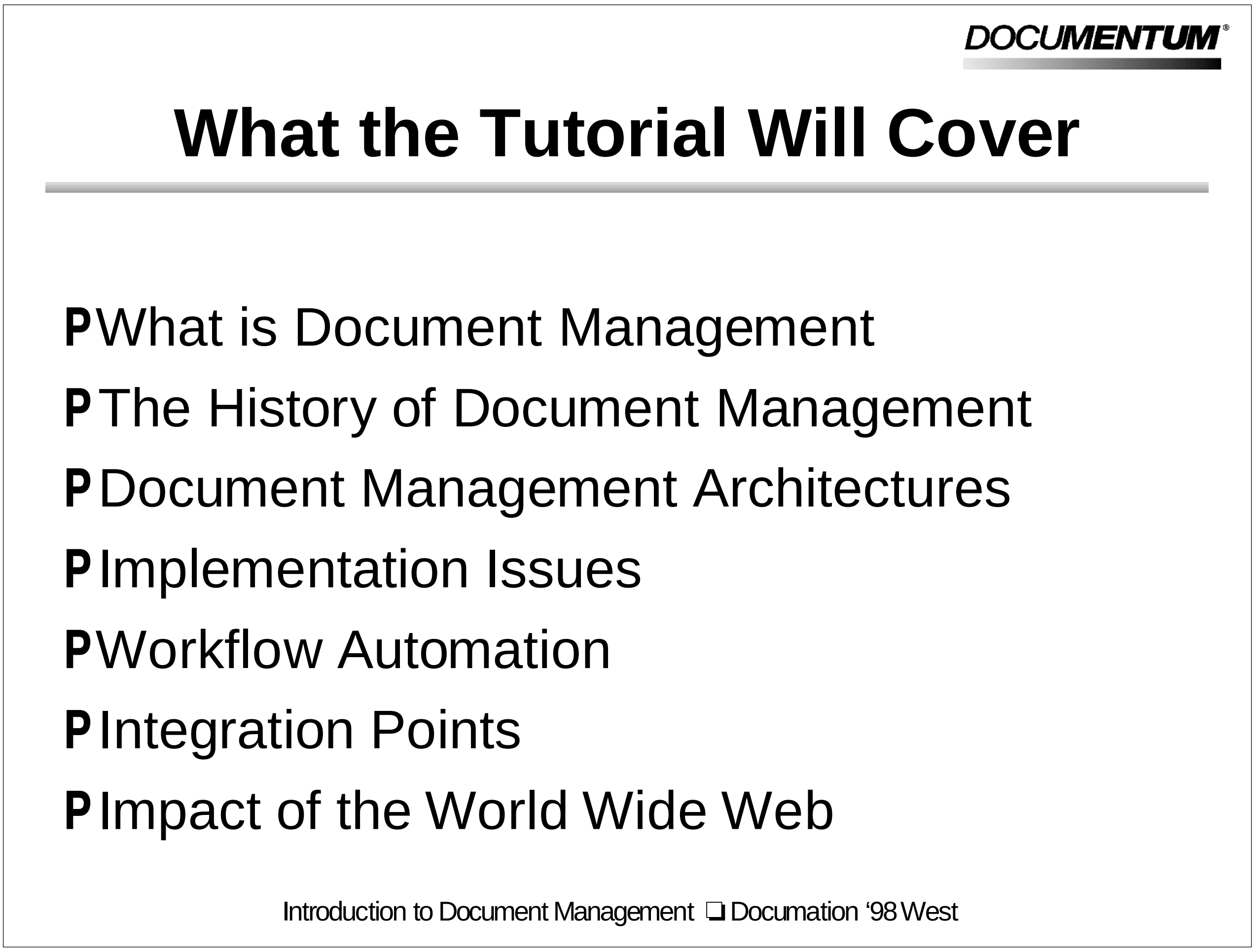
Documents as Strategic Assets

P Critical to complex organizational behaviors
- ▸ Provide context
- ▸ Integrate, document, and communicate understanding

P Critical to market success
- ▸ Product utilization
- ▸ Customer satisfaction

P Inconsistently recognized as strategic
- ▸ Real men do databases
- ▸ CALS, ATA 2000, ISO 9000, etc.

Introduction to Document Management ❑ Documation ‘98 West



DOCUMENTUM®

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

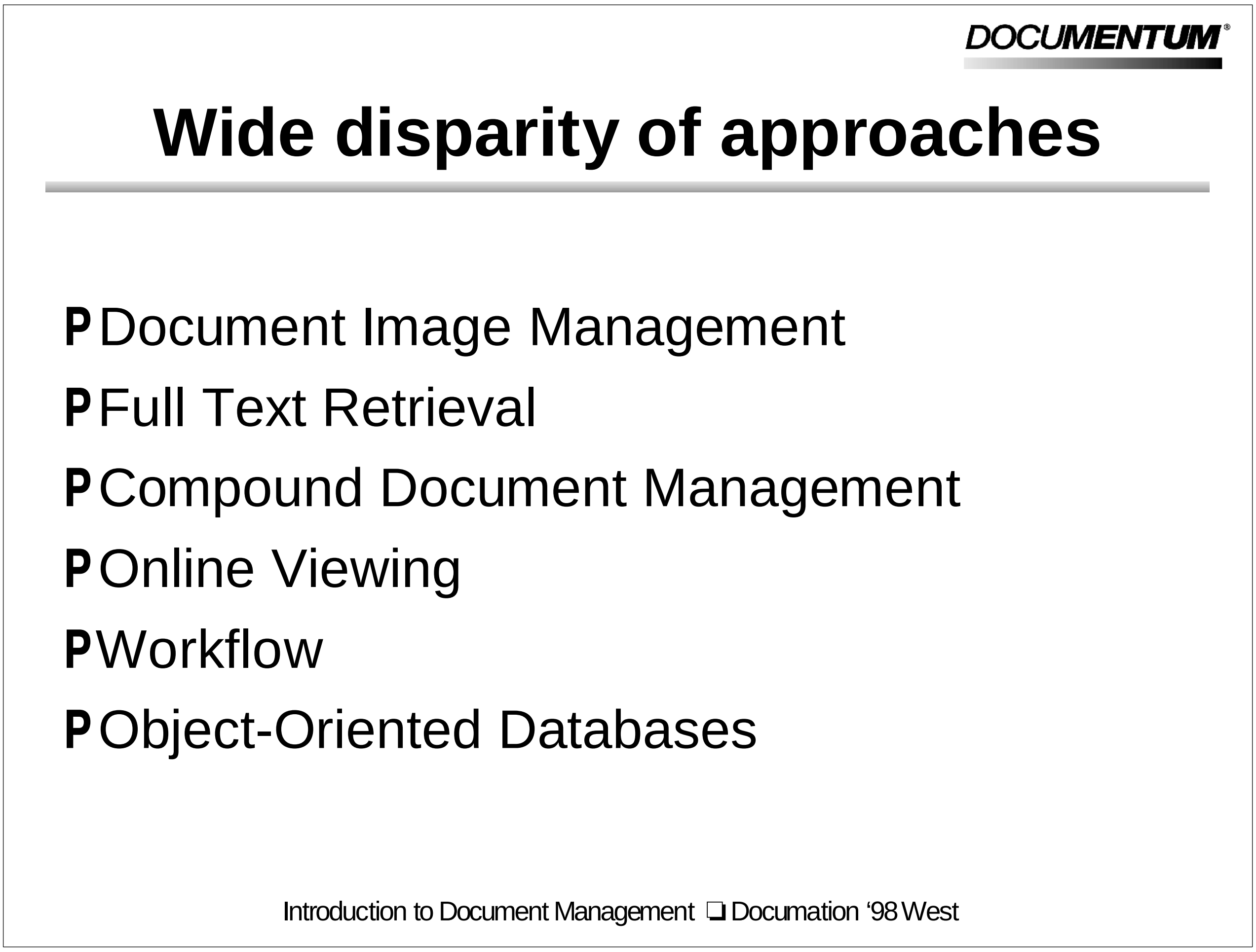
P Impact of the World Wide Web

Introduction to Document Management ❑ Documation ‘98 West

What is Document Management?

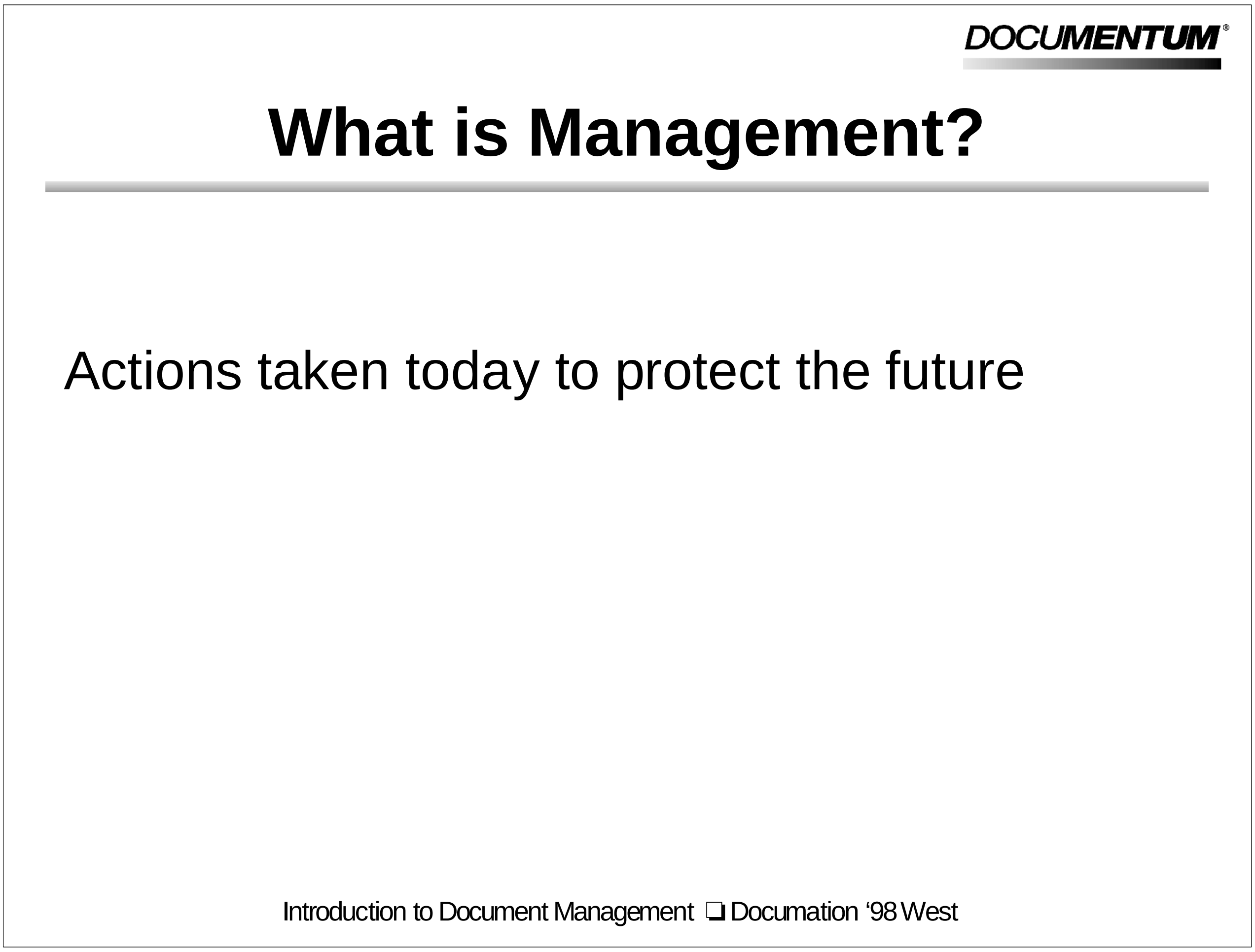
Simple Definition

Systems for managing collections of documents



Wide disparity of approaches

- P Document Image Management
- P Full Text Retrieval
- P Compound Document Management
- P Online Viewing
- P Workflow
- P 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 = f(Legacy, Vision)

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

Metadata Determines Future Value

P Metadata = data about data

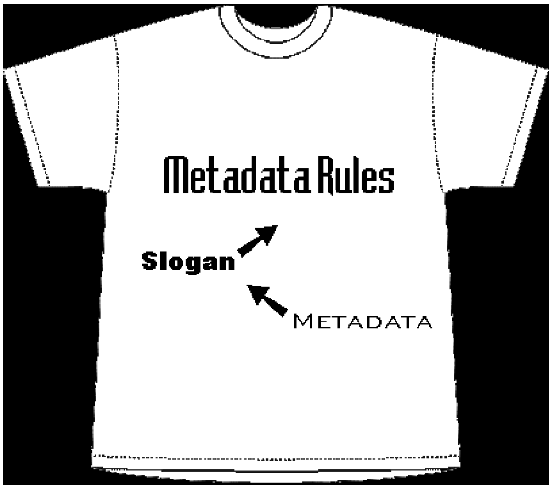
P Metadata is the basis for behavior

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
  - ‣ Emphasize different definitions of document value
  - ‣ Are tied to different classes of metadata

History of Document Management Systems

History Overview

P Mirrors the evolving concept of the document

P Tied to technology and metadata changes
- Chicken and egg
- Organizational learning
- Behavioral implications

P Four stages
- Paper documents
- Automated paper documents
- Electronic documents
- Active documents

Paper Documents

Behavioral focus: 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

Behavioral focus: Generating paper documents

- P Metadata emphasizes visual formatting
- P Laser printers allow more addressability and control
- P Tools function like fast, powerful pens
- P Metadata / operator interaction based on formatting codes (procedural markup)

Automated Paper Documents

Performance criteria

- P Personal productivity
- P Visual sophistication
- P Speeding revisions to paper documents
- P Lifecycle costs de-emphasized
  - ▸ Hidden costs
  - ▸ Diseconomies ("info pollution")
- P Need for interchange drives adoption of standardized encodings

Automated Paper Documents

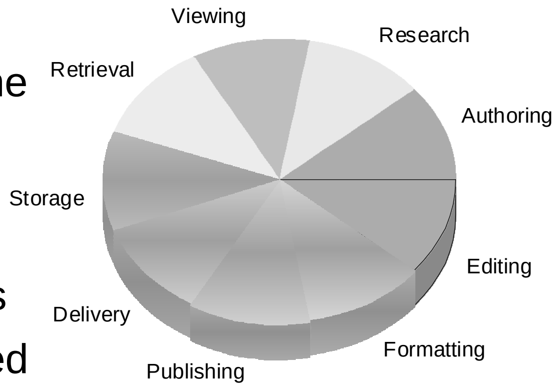
Management systems and supporting technologies

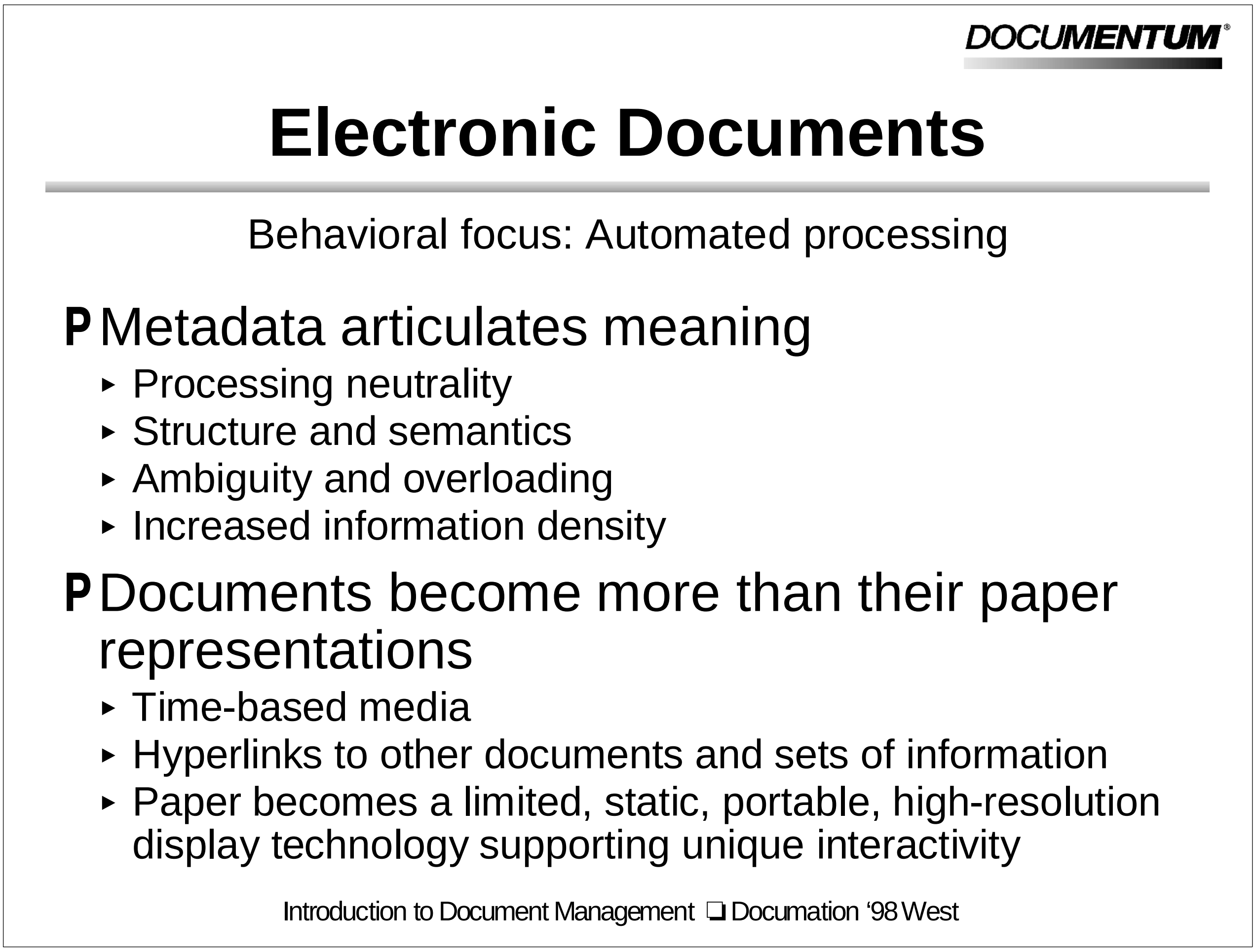
- P Manage information *about* the documents
  - ▸ File management systems
  - ▸ Image management systems
  - ▸ Other database-based indexing systems
- P Manipulate document appearance
  - ▸ Graphics, wordprocessing, and desktop publishing tools
- P Management of meaning and semantics limited to relational database world

Automated Paper Documents

Limitation: Illusion of control

- P Paper hides a multitude of sins
- P Solutions focus on a subset of the document lifecycle
- P Personal productivity drives suboptimization
- P Paper-based interface standards
- P Human interpretation still required
- P Limited capacity for automated conversions and transformations

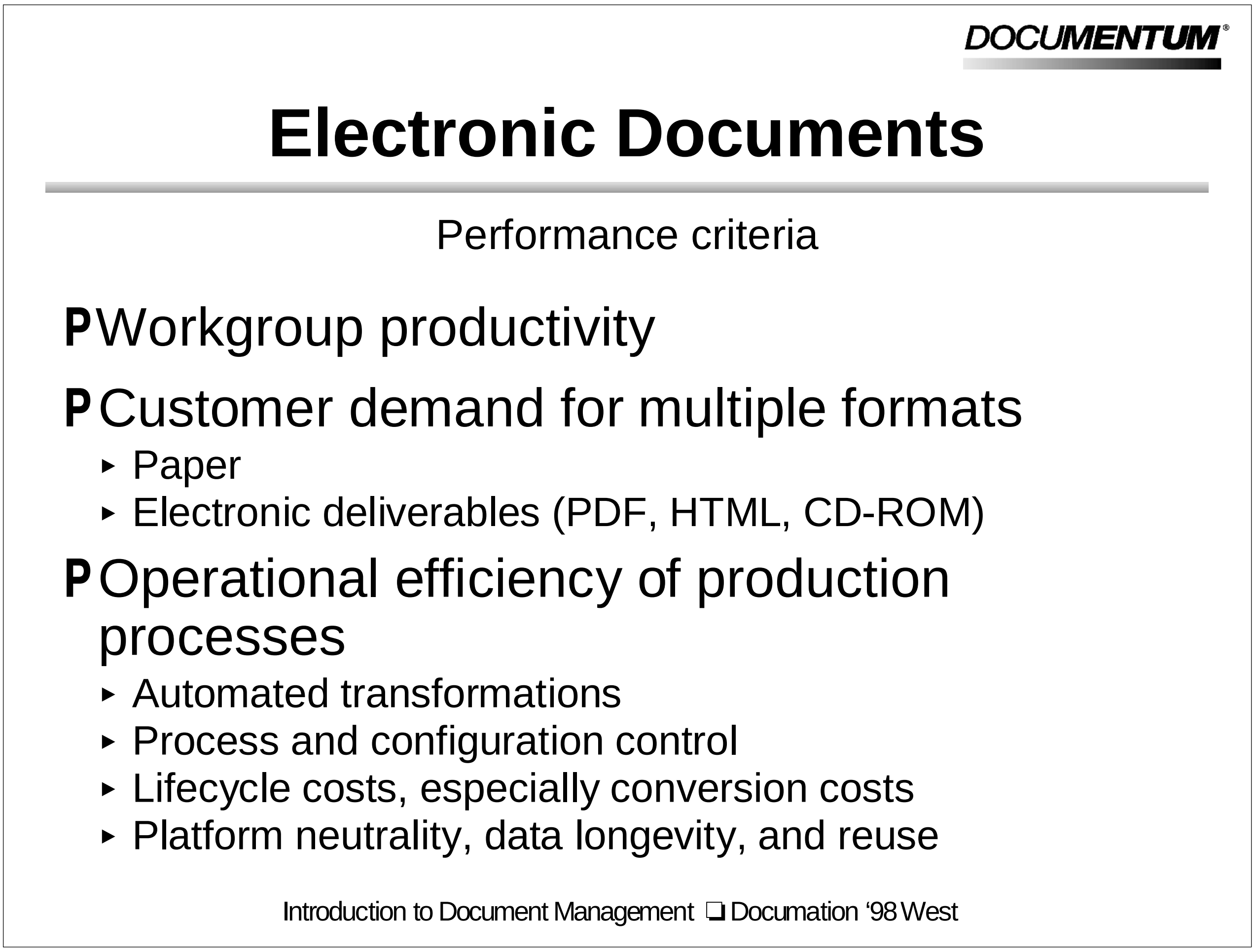




Electronic Documents

Behavioral focus: Automated processing

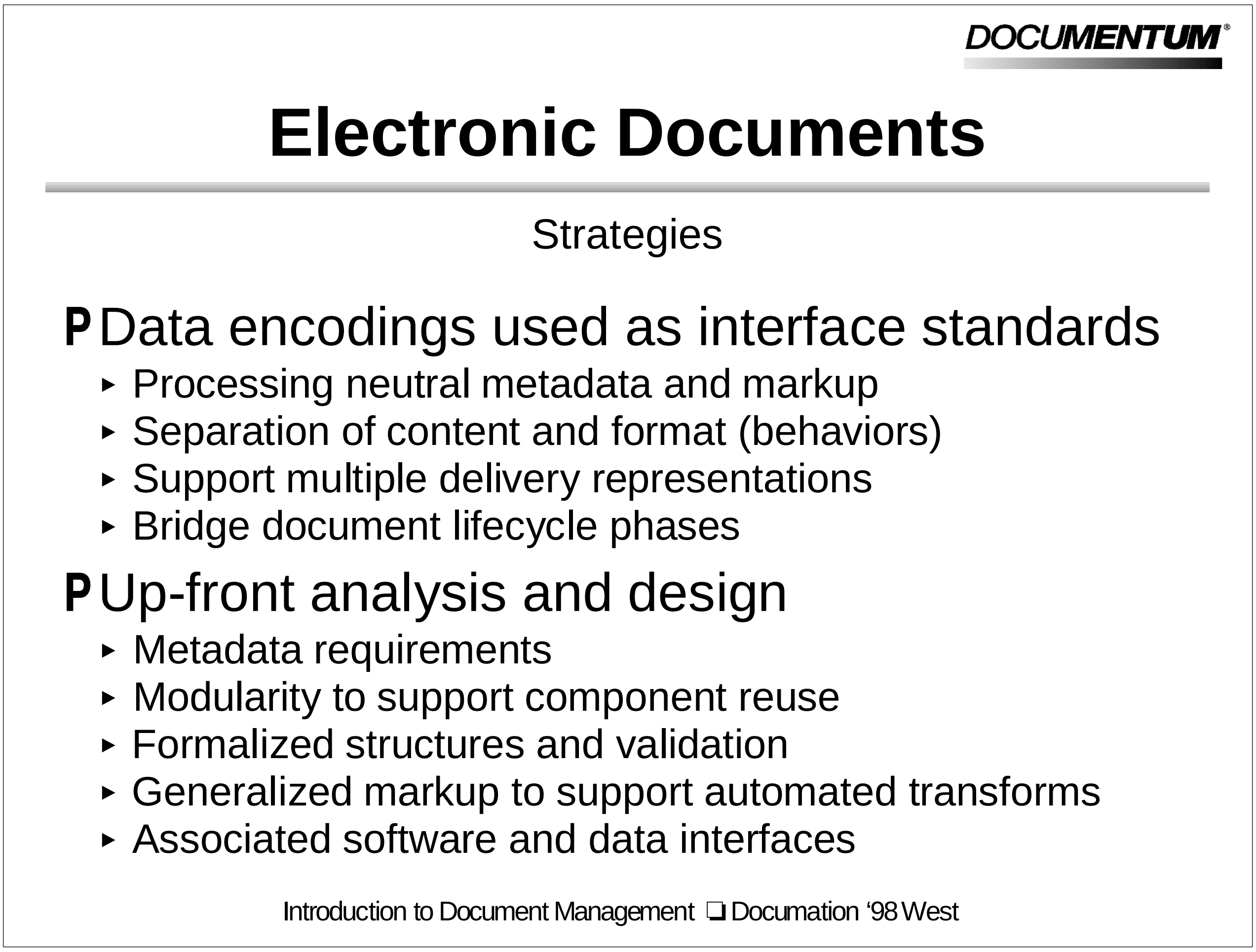
- P Metadata articulates meaning
  - Processing neutrality
  - Structure and semantics
  - Ambiguity and overloading
  - Increased information density
- P Documents become more than their paper representations
  - Time-based media
  - Hyperlinks to other documents and sets of information
  - Paper becomes a limited, static, portable, high-resolution display technology supporting unique interactivity



Electronic Documents

Performance criteria

- P Workgroup productivity
- P Customer demand for multiple formats
  - Paper
  - Electronic deliverables (PDF, HTML, CD-ROM)
- P Operational efficiency of production processes
  - Automated transformations
  - Process and configuration control
  - Lifecycle costs, especially conversion costs
  - Platform neutrality, data longevity, and reuse



DOCUMENTUM®

Electronic Documents

Strategies

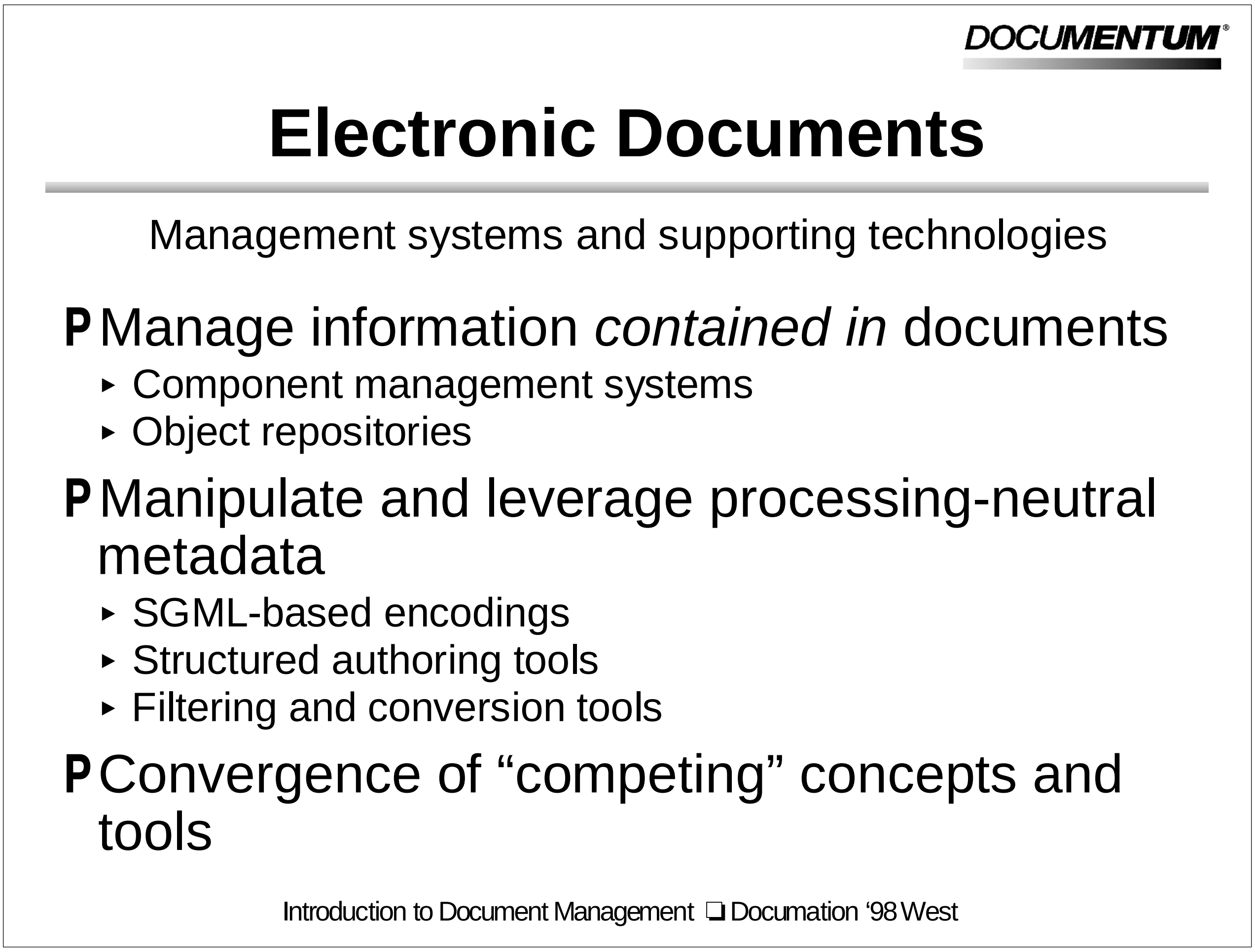
P Data encodings used as interface standards

- Processing neutral metadata and markup
- Separation of content and format (behaviors)
- Support multiple delivery representations
- Bridge document lifecycle phases

P Up-front analysis and design

- Metadata requirements
- Modularity to support component reuse
- Formalized structures and validation
- Generalized markup to support automated transforms
- Associated software and data interfaces

Introduction to Document Management ❑ Documation ‘98 West



DOCUMENTUM®

Electronic Documents

Management systems and supporting technologies

P Manage information *contained in* documents

- Component management systems
- Object repositories

P Manipulate and leverage processing-neutral metadata

- SGML-based encodings
- Structured authoring tools
- Filtering and conversion tools

P Convergence of “competing” concepts and tools

Introduction to Document Management ❑ Documation ‘98 West

Electronic Documents

Issues

P Shared pools and information reuse drive new organizational models

- ▸ Operational performance and downstream value increasingly limited by authoring process
- ▸ Synchronization of process and technology changes

P Automation drives standardized designs

P Politics, authority, and autonomy

Electronic Documents

Limitations: Infoglut and accessability

P Improved production processes drive infoglut

- ▸ Availability drives inability to access and leverage
- ▸ So much information, so few answers
- ▸ Metadata quality and relevance a critical limiting factor

P Information access

- ▸ Retrieval: precision and recall
- ▸ Critical information often lies at the intersection points of complex, multi-dimensional conceptual frameworks
- ▸ Combinatorial explosion of retrieval criteria

Electronic Documents

Limitation: Access does not guarantee performance

P Documents are valuable because they provide context to information

P Embedded metadata often inappropriate and/or irrelevant to new behavioral domains

- Topic
- Wording and tone
- Pedigree
- Rationale
- Conceptual framework

Active Documents

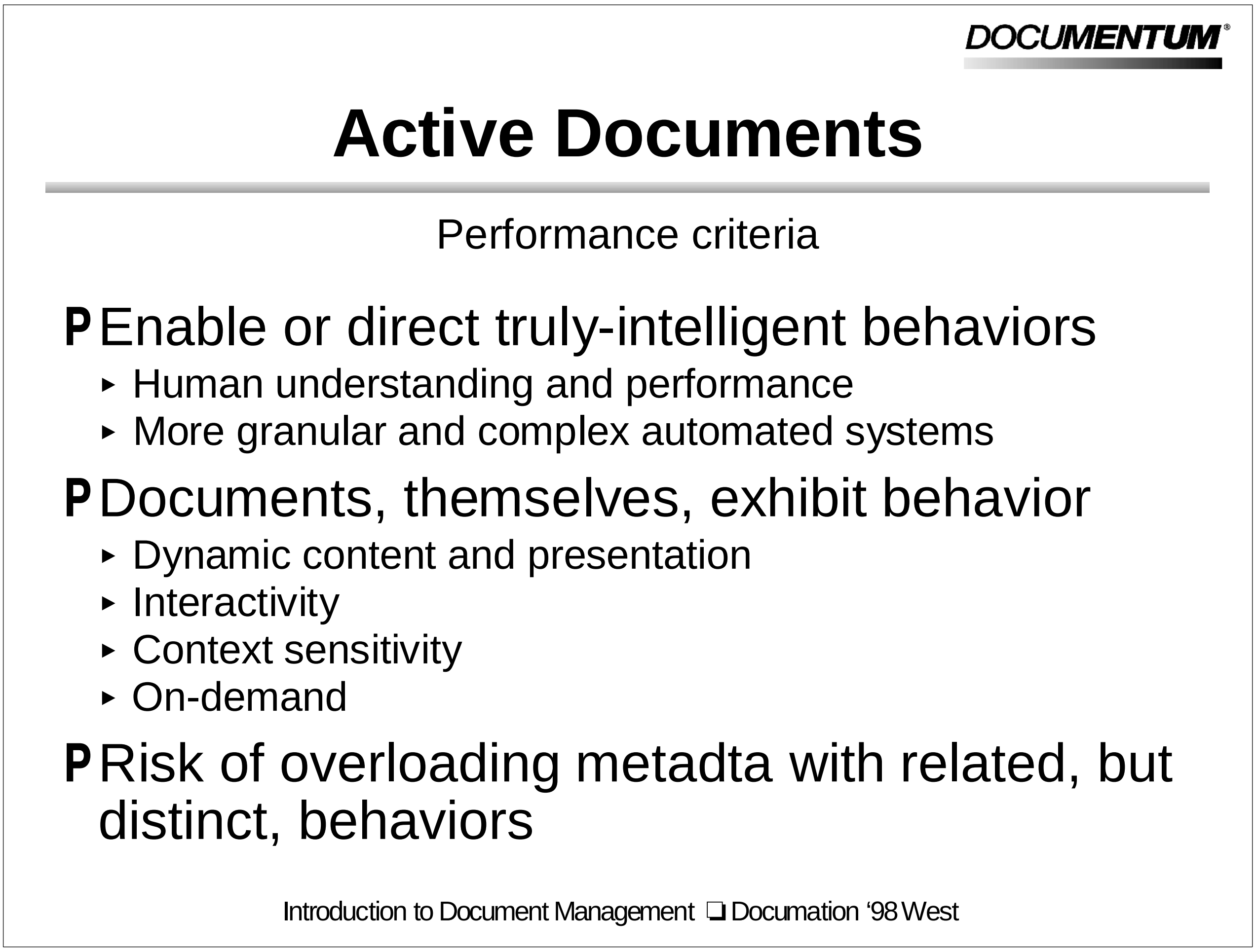
Behavioral focus: Integration of related systems

P Metadata

- Codifies behavioral value of information in documents
- Supports multiple behavioral domains

P Documents shift from static artifacts to dynamic views

- Transient and more short-lived
- Query-based assembly
- Conditionality and effectivity
- Relevant views reflect the intersection of multiple criteria



DOCUMENTUM®

Active Documents

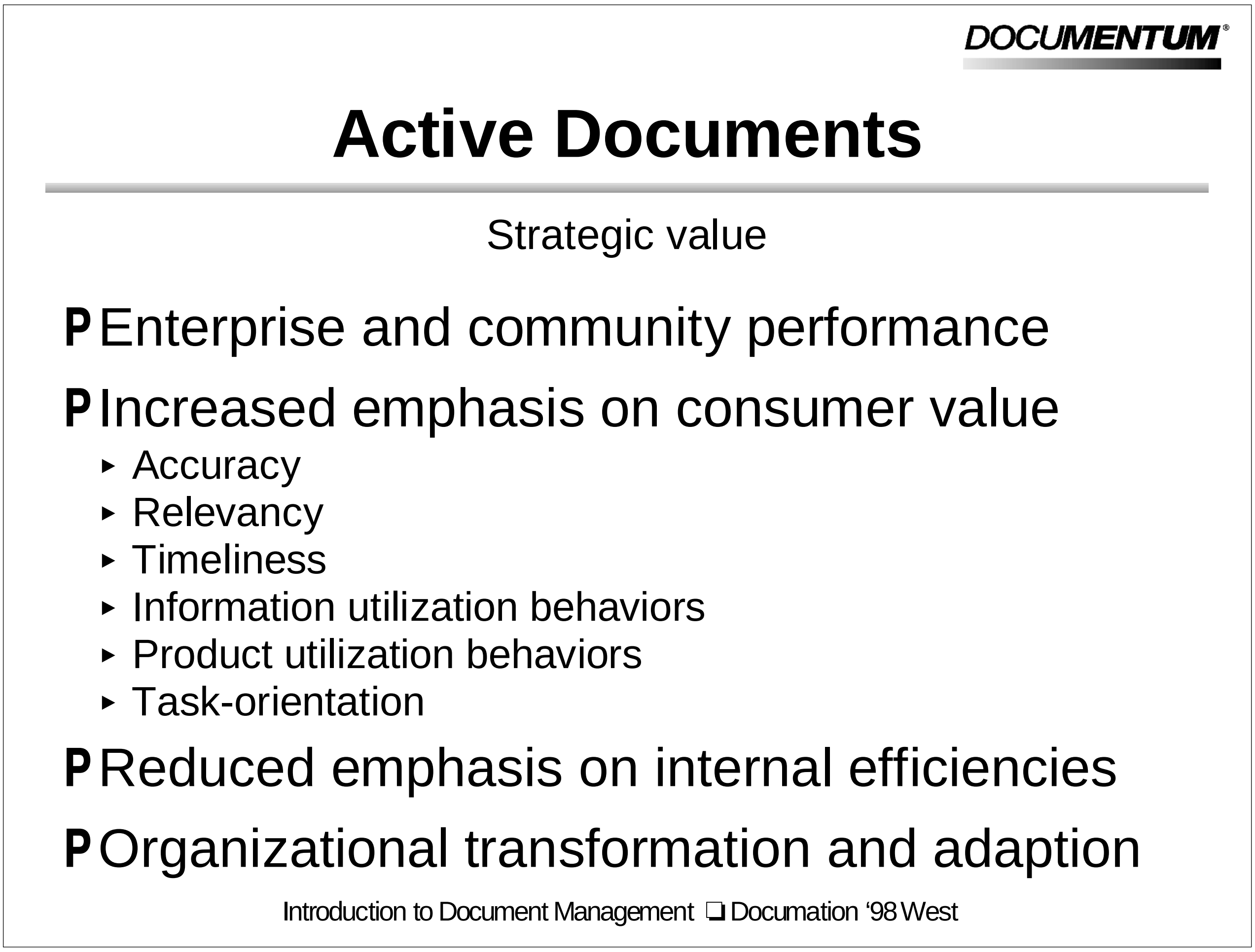
Performance criteria

P Enable or direct truly-intelligent behaviors
- ‣ Human understanding and performance
- ‣ More granular and complex automated systems

P Documents, themselves, exhibit behavior
- ‣ Dynamic content and presentation
- ‣ Interactivity
- ‣ Context sensitivity
- ‣ On-demand

P Risk of overloading metadta with related, but distinct, behaviors

Introduction to Document Management ❑ Documation '98 West



DOCUMENTUM®

Active Documents

Strategic value

P Enterprise and community performance

P Increased emphasis on consumer value
- ‣ Accuracy
- ‣ Relevancy
- ‣ Timeliness
- ‣ Information utilization behaviors
- ‣ Product utilization behaviors
- ‣ Task-orientation

P Reduced emphasis on internal efficiencies

P Organizational transformation and adaption

Introduction to Document Management ❑ Documation '98 West

Active Documents

Management strategies

P Focus on the behavioral implications of documents (knowledge utilization events)

P Integrate the entire document lifecycle and associated knowledge lifecycles

P Shift from engineered to organic systems and organizations

- ▸ Decentralization
- ▸ Distributed and autonomous decision making
- ▸ Multiple goals
- ▸ Disequilibrium

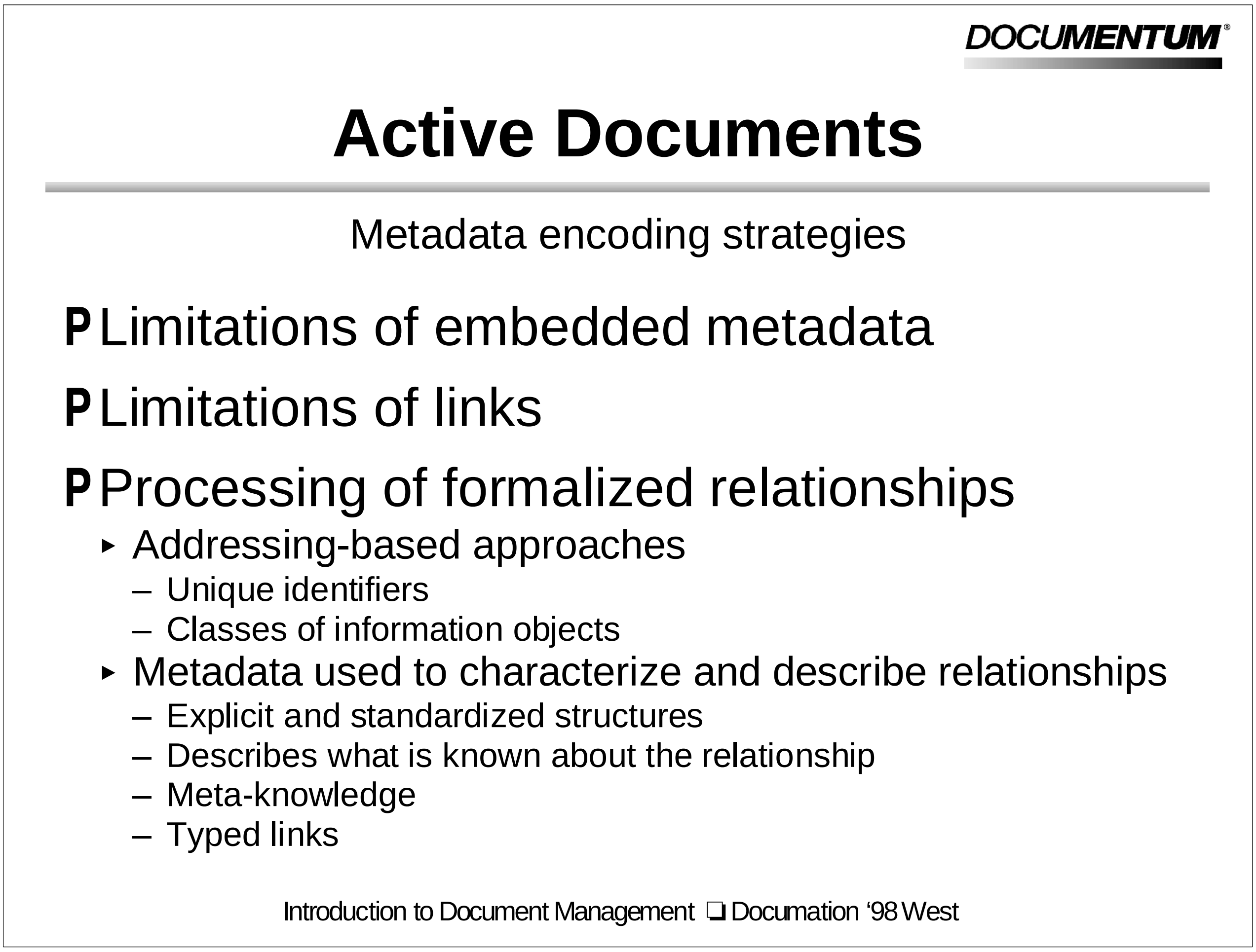
Active Documents

Multidimensional metadata strategies

P Generalized descriptions

P Process-specific descriptions

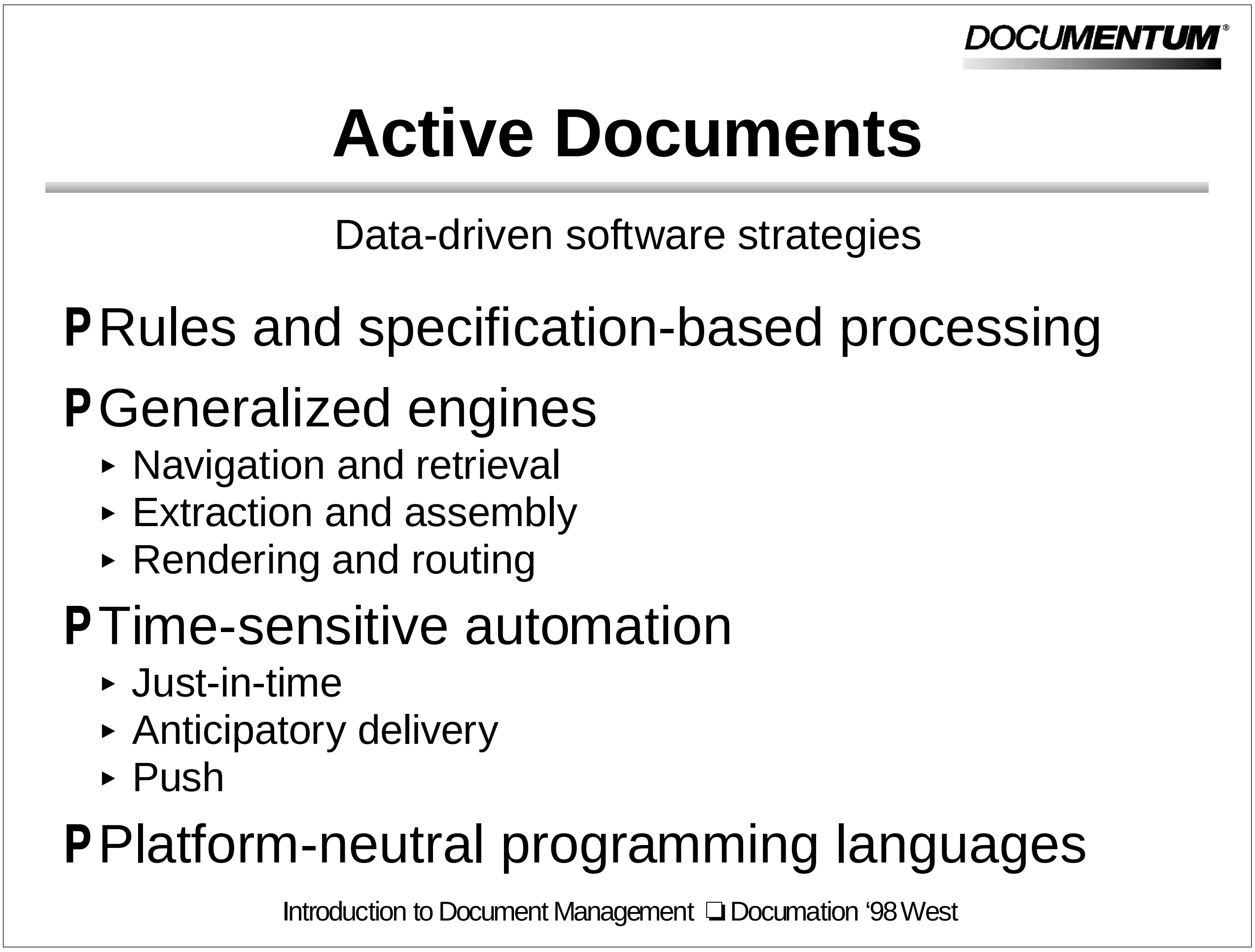
- ▸ Audience profiles
- ▸ Models of human behavior
- ▸ Models of technical systems and behaviors
- ▸ Transformations
  - – Current state
  - – Past and future state changes and transformations
  - – Pedigree
  - – Rationale
- ▸ Other behavioral and conceptual domains



Active Documents

Metadata encoding strategies

- P Limitations of embedded metadata
- P Limitations of links
- P Processing of formalized relationships
  - ▸ Addressing-based approaches
    - – Unique identifiers
    - – Classes of information objects
  - ▸ Metadata used to characterize and describe relationships
    - – Explicit and standardized structures
    - – Describes what is known about the relationship
    - – Meta-knowledge
    - – Typed links



Active Documents

Data-driven software strategies

- P Rules and specification-based processing
- P Generalized engines
  - ▸ Navigation and retrieval
  - ▸ Extraction and assembly
  - ▸ Rendering and routing
- P Time-sensitive automation
  - ▸ Just-in-time
  - ▸ Anticipatory delivery
  - ▸ Push
- P Platform-neutral programming languages

Active Documents

Management systems and supporting technologies

P Manage the *relationships* described by and associated with documents

- ▸ Document fragments (increased granularity)
- ▸ Behavioral fragments
- ▸ Non-linear and intersecting revisions
- ▸ Version clusters

P Direct, track, and record multiple behaviors

- ▸ Hypermedia authoring (links and annotations)
- ▸ Temporal processing (workflow)
- ▸ Transformations (stylesheets, conversions, mappings)

What is Document Management?

Revisited

P Today's high-performance documents are based on meanings and relationships

P Emphasis is shifting away from

- ▸ Simple storage and retrieval
- ▸ Independent management of life cycle phases

P New emphasis on integrating interrelated information and knowledge lifecycles

P Systems often encompass competing concepts of the document

DOCUMENTUM®

Overview of Document Management Architectures

DOCUMENTUM®

Overview

P Four models

- Image-based
- WYSIWYG DTP
- Compound document management
- Knowledge management / multidimensional relationship management

P Components

- Data encoding standards
- Software interoperability standards
- Task-specific tools
- Communications and repository infrastructure

Image-based Architectures

P Dragging paper documents into the electronic age

P Heavy reliance on human interpretation

P Layering of metadata to capture meaning and understanding

P Workflow automation and annotation innovations

WYSIWYG DTP

P Control of visual aspects

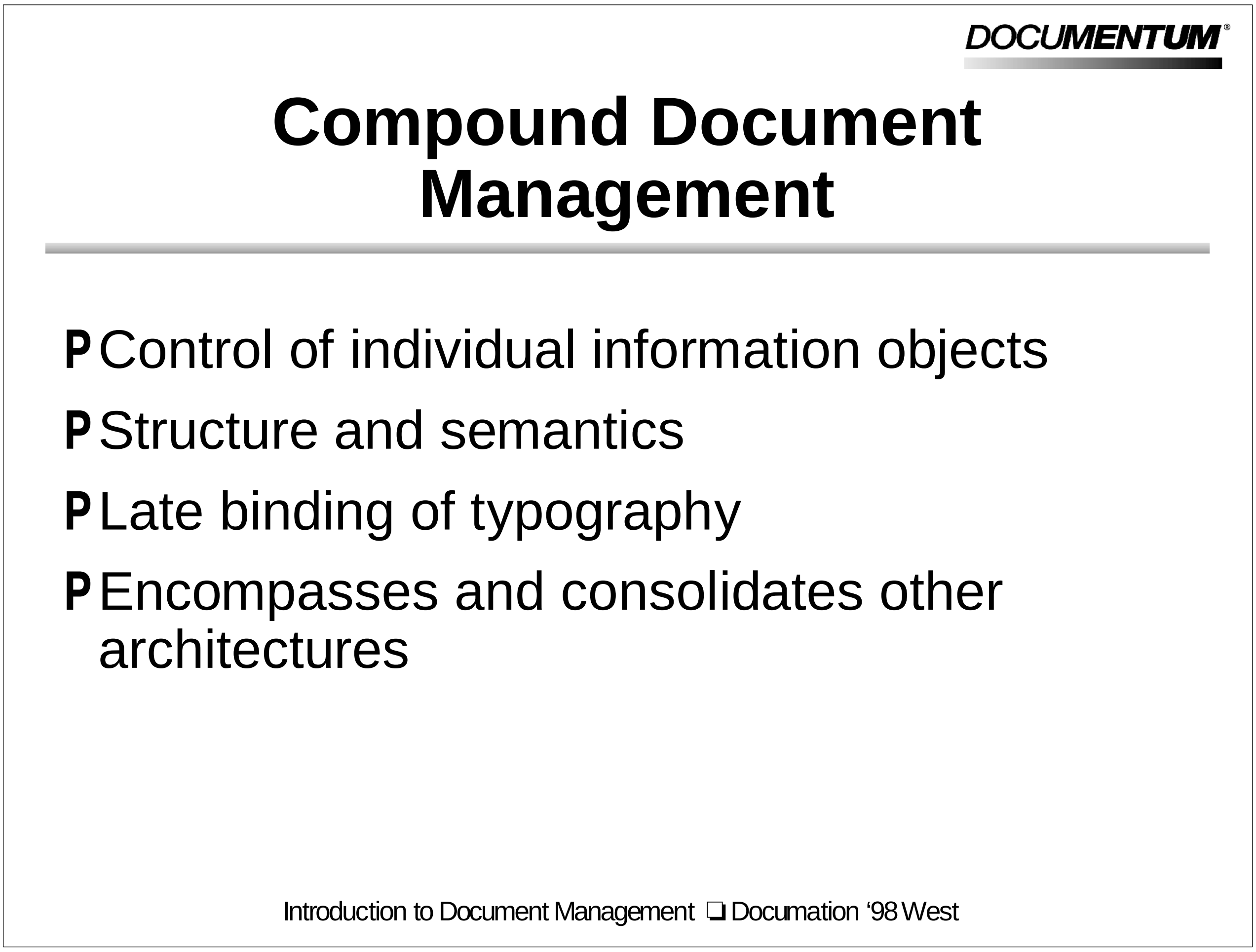
P File-based and BLOBS

P Production focus

P Short-lived documents

- ‣ Advertising
- ‣ Novelty
- ‣ Drama

P WWW



DOCUMENTUM®

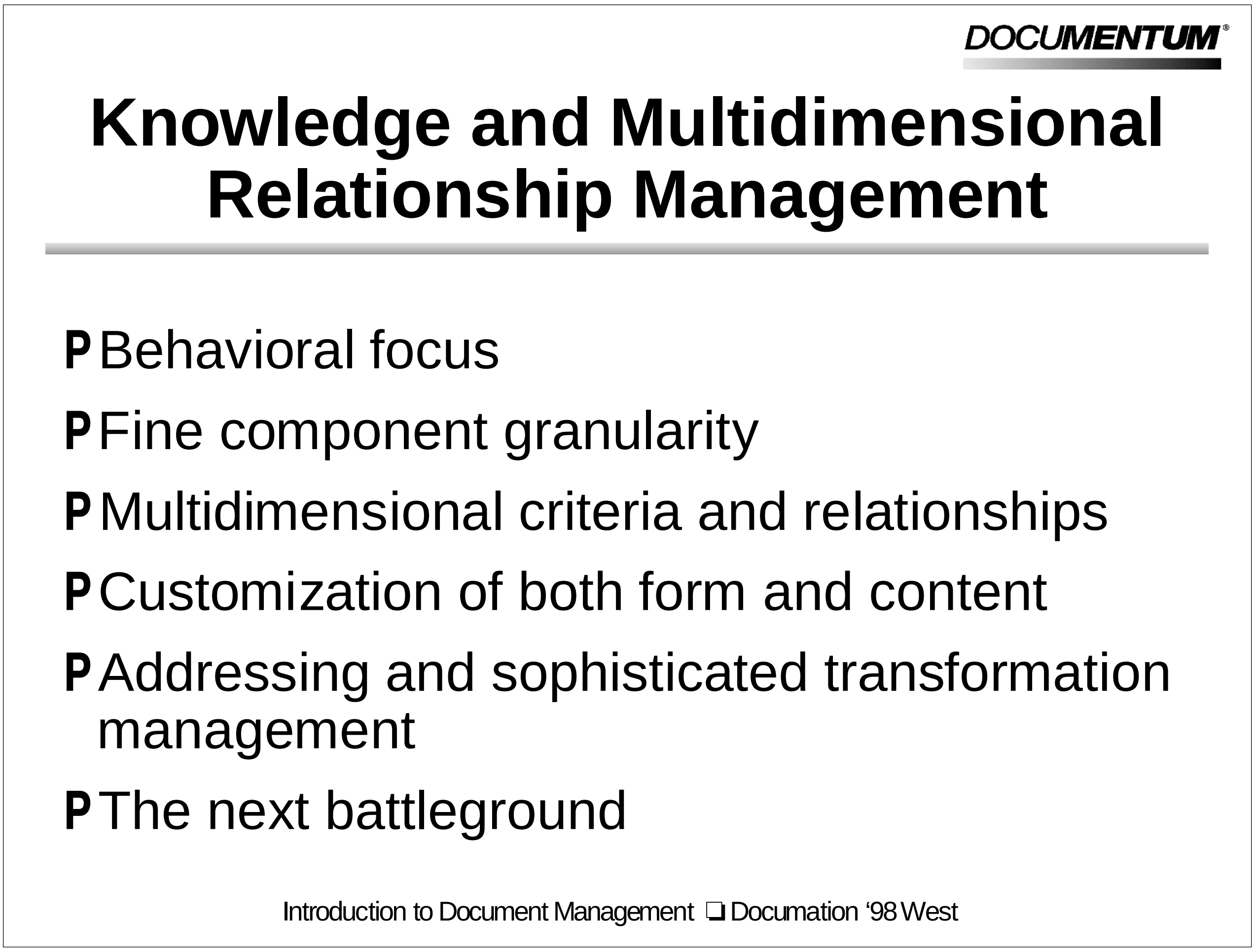
Compound Document Management

P Control of individual information objects

P Structure and semantics

P Late binding of typography

P Encompasses and consolidates other architectures



DOCUMENTUM®

Knowledge and Multidimensional Relationship Management

P Behavioral focus

P Fine component granularity

P Multidimensional criteria and relationships

P Customization of both form and content

P Addressing and sophisticated transformation management

P The next battleground

Data Encoding Standards

General Questions

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

Data Encoding Standards

Text

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

Data Encoding Standards

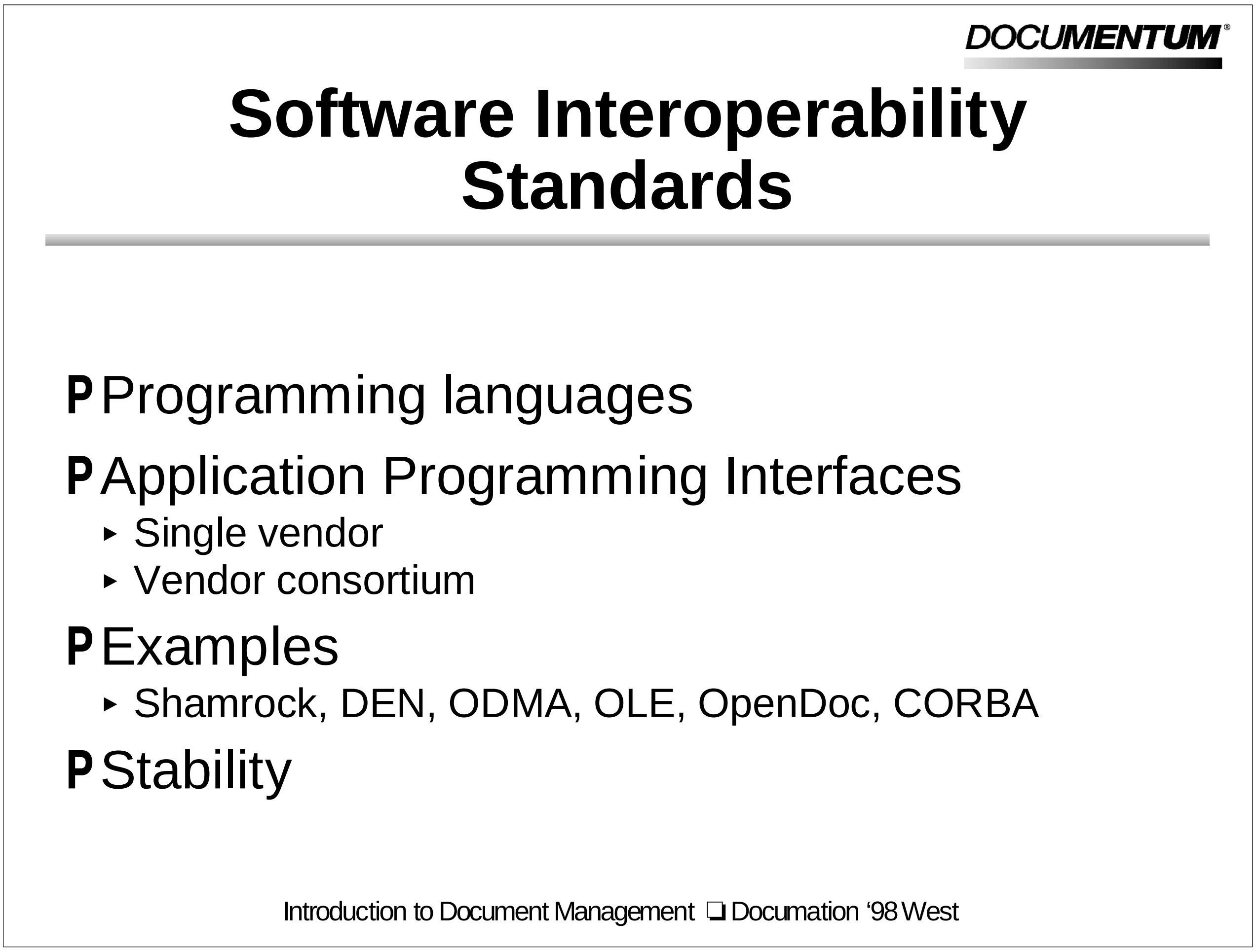
Graphics

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

Data Encoding Standards

Other

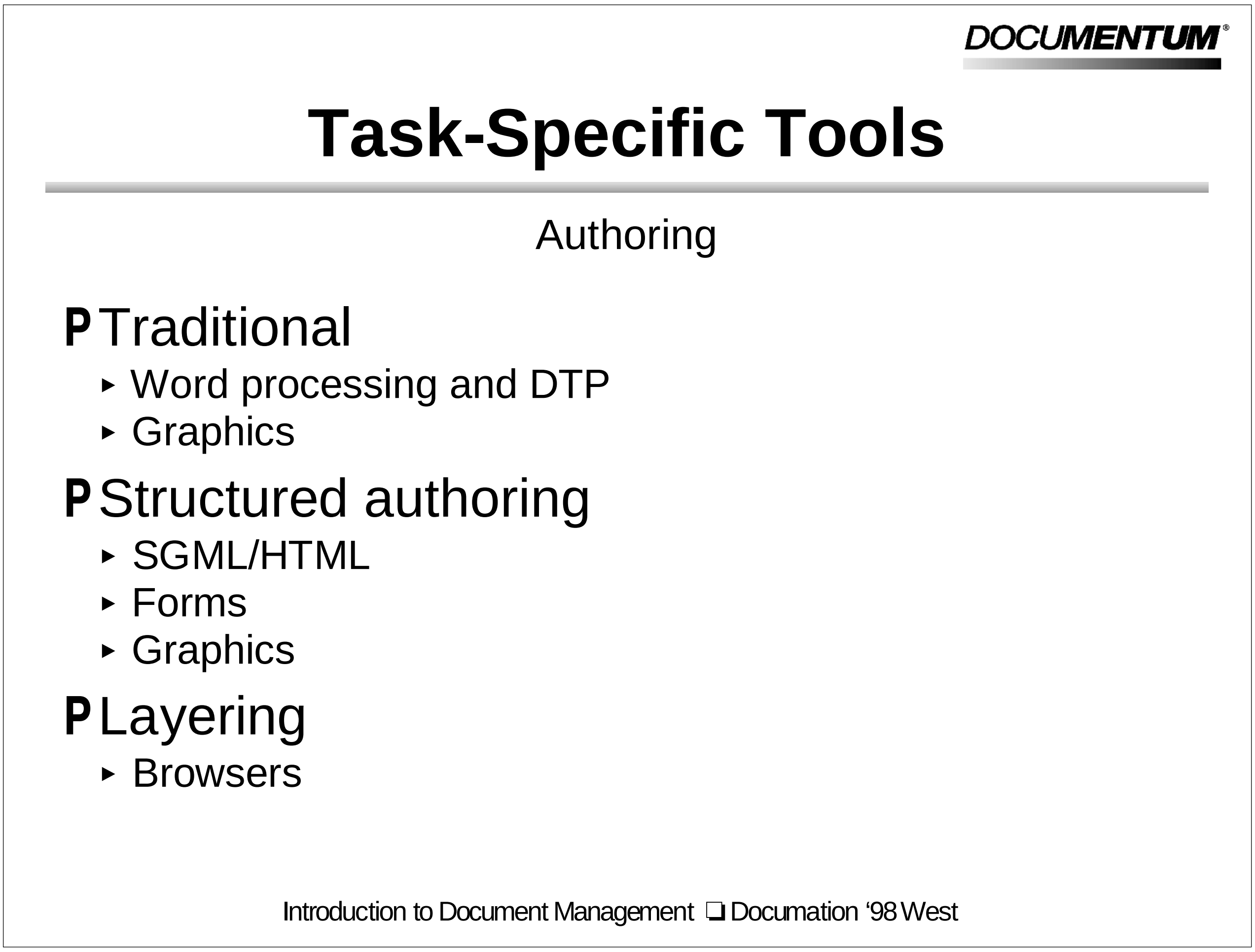
- P Audio
- P Video
- P Voice
- P Positional / GIS
- P Hyperlinking
- P Rendering
- P Behaviors



DOCUMENTUM®

Software Interoperability Standards

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



DOCUMENTUM®

Task-Specific Tools

Authoring

- P Traditional
  - ‣ Word processing and DTP
  - ‣ Graphics
- P Structured authoring
  - ‣ SGML/HTML
  - ‣ Forms
  - ‣ Graphics
- P Layering
  - ‣ Browsers

Task-Specific Tools

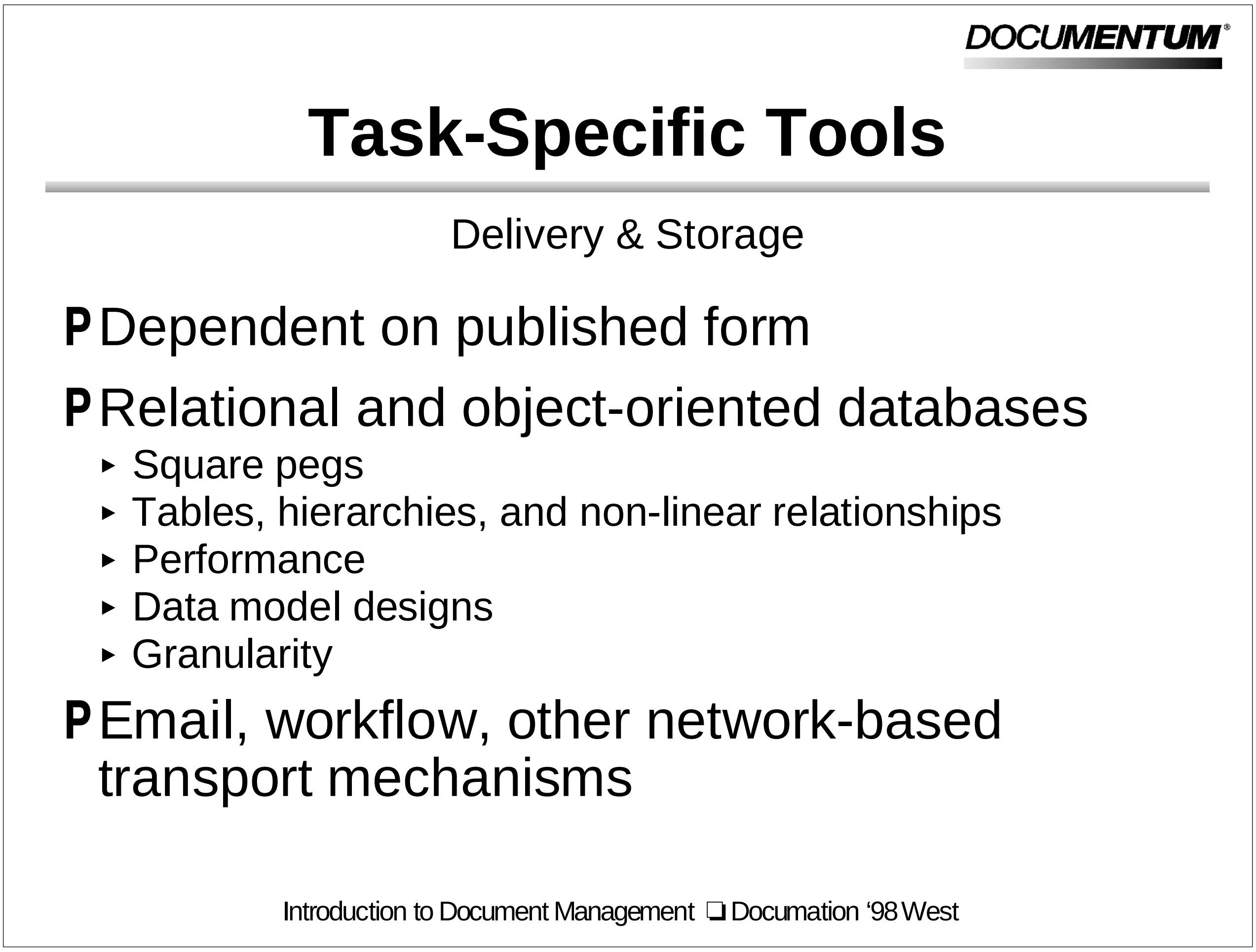
Editing

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

Task-Specific Tools

Formatting & Publishing

- Converters
  - Scanners
  - OCR/vectorizers
  - Programmable
- Composition tools
- Physical media and associated hardware
- Hypermedia authoring tools
- Print on demand



Task-Specific Tools

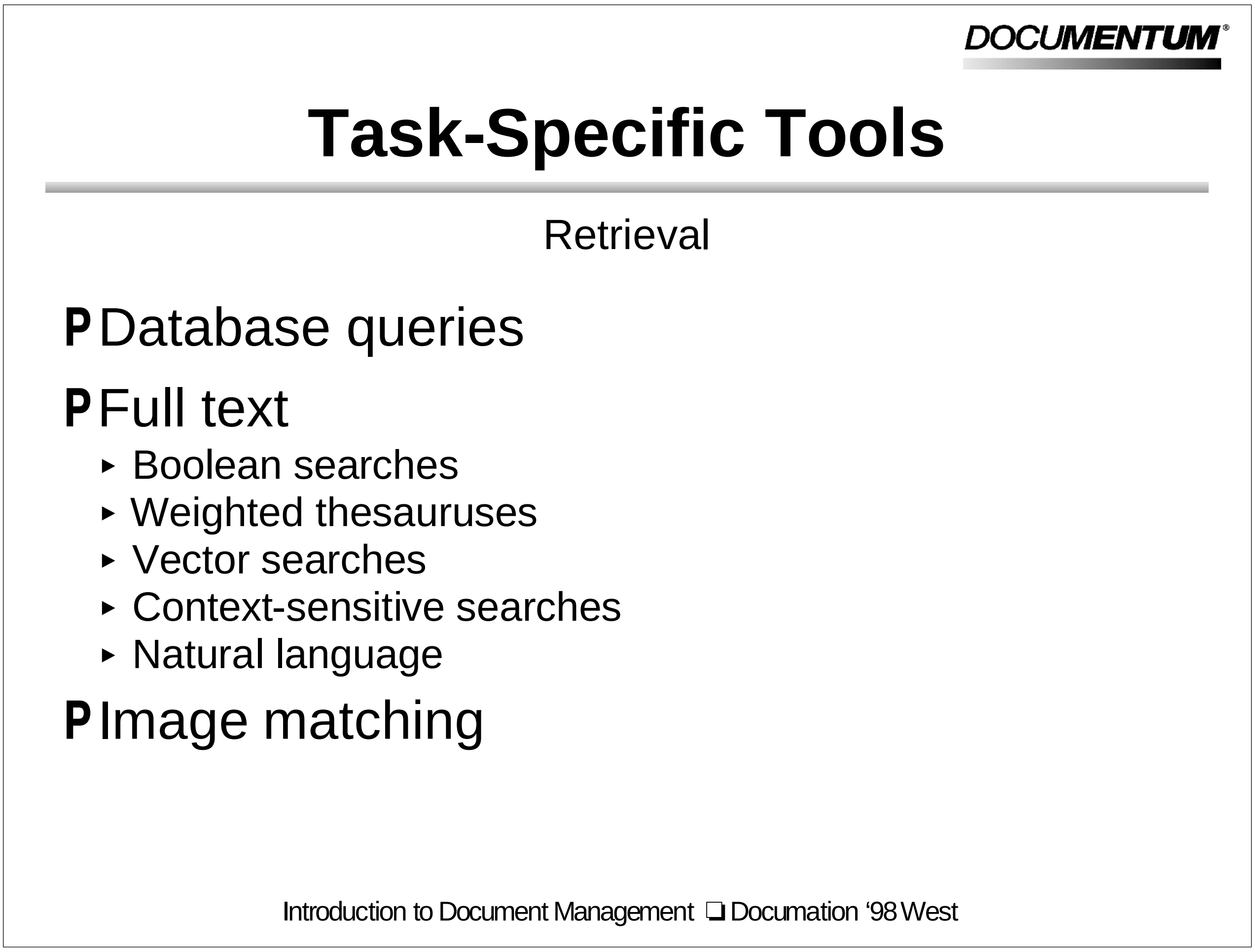
Delivery & Storage

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

P Text readers

P Native file viewers

P Raster viewers

P Page viewers

P Binary browsers

P Fixed markup language browsers

P Arbitrary DTD browsers

Infrastructure

P Repository and communications subsystems

P Scope

P Granularity

P Encodings

P Versioning and configuration control

P Target of most software interoperability standards

Implementation Issues

Human Issues

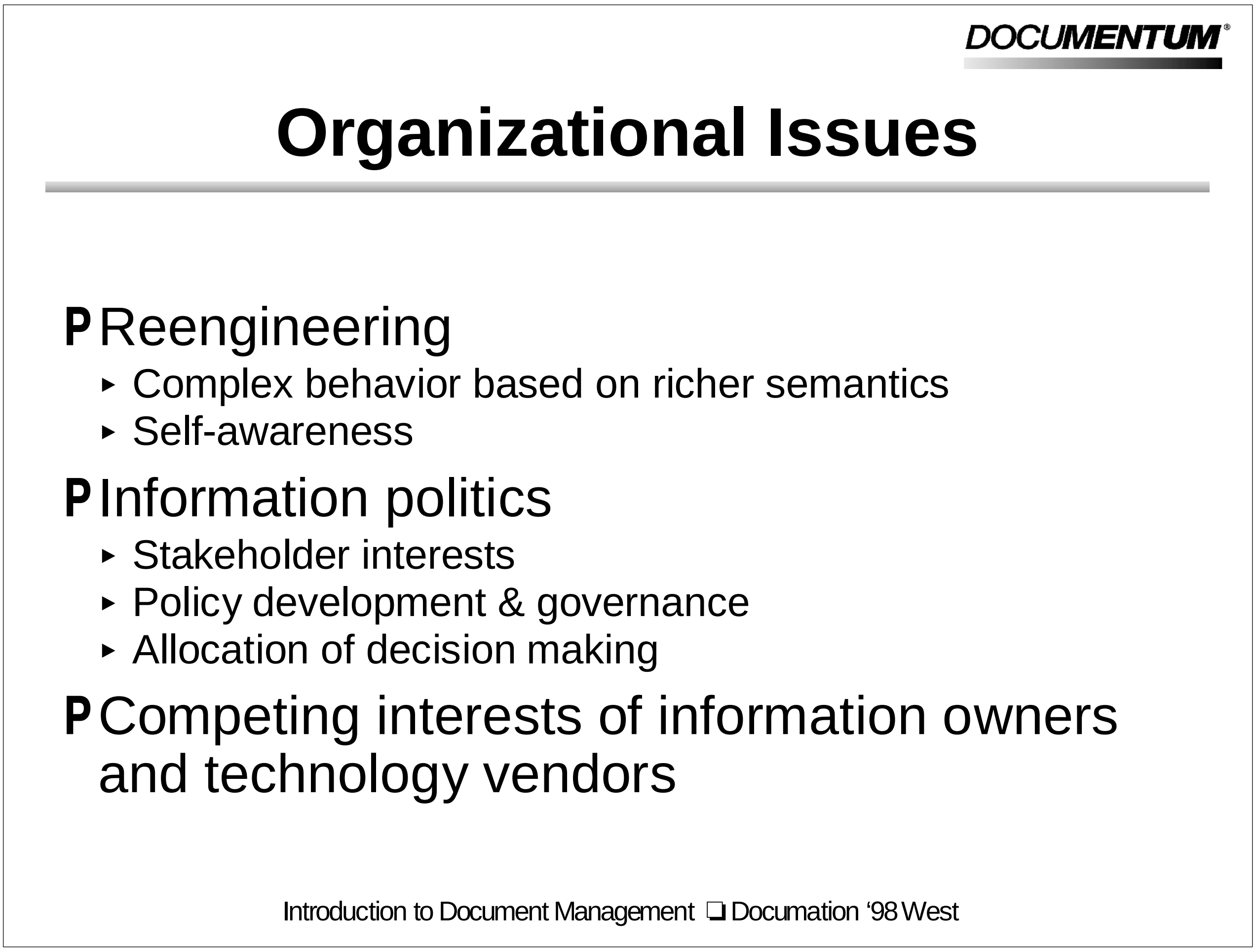
P Difficulty of adopting enabling technologies

- ▸ Conceptualization
- ▸ Learning
- ▸ Foresight

P Perceptions

- ▸ Technology problem
- ▸ Uniqueness

P Who knows?



DOCUMENTUM®

Organizational Issues

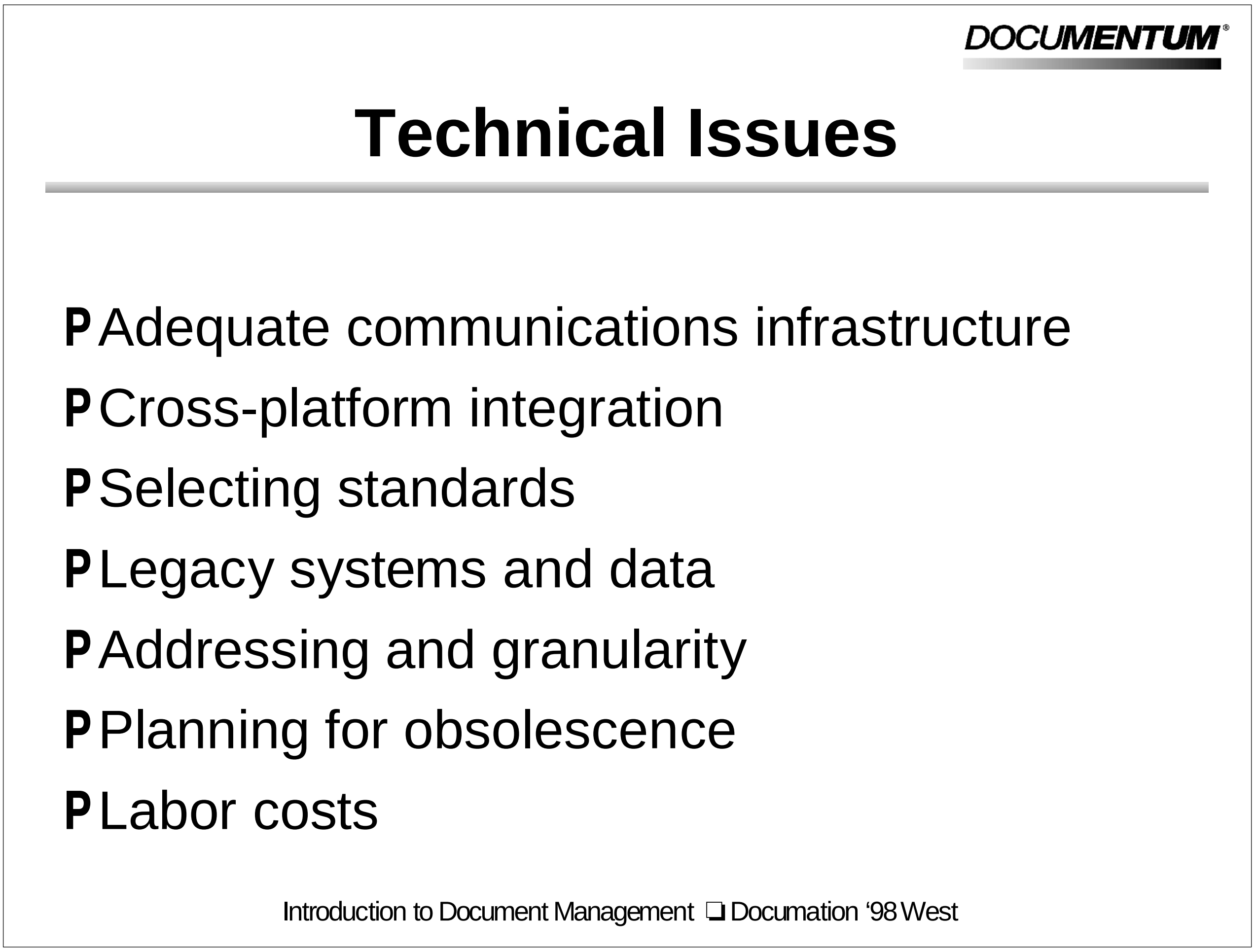
P Reengineering

- ‣ 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



DOCUMENTUM®

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

P Often confused with document management
- ▸ Check-in and check-out
- ▸ Rules-based processing
- ▸ vs component-level configuration control

P Convergence with document management
- ▸ Routing and communication

P 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

P Information suppliers and consumers

P Metadata requirements

P Process, policy, politics

P 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

- P Focus for consolidation
- P Encoding standards
  - ‣ HTML
  - ‣ XML
- P Software design

Focus for Consolidation

- P Aim for the accident
- P Change changes change
  - ‣ Perceptions of value
  - ‣ User needs
  - ‣ Vendor desires
  - ‣ Laboratory for innovation

HTML

P HTML hides a multitude of sins

P A application of SGML

- ▸ Tagset history
- ▸ Conformance issues
- ▸ Volatility
- ▸ Theology

P Easy to get into

P Danger in thinking that more than a delivery encoding

HTML

Issues and strategies

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

Extensible Markup Language

- P Drivers
  - ▸ Browser wars
  - ▸ Industrial requirements
- P SGML application profile
  - ▸ Conformance to ISO standard
  - ▸ Reduced feature set
  - ▸ Well-formed documents
- P Companion standards
  - ▸ Extensible Style Language (XSL)
  - ▸ Extensible Linking Language (XLL)

Extensible Markup Language

Market impacts

- P Bridging
  - ▸ SGML and application development communities
  - ▸ Document management and financial services
- P New baseline for relationship management
  - ▸ Codify demand
  - ▸ Define technical standards
- P Destabilize
  - ▸ Tools
  - ▸ Interfaces
  - ▸ Market segmentation

Software Design

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

Conclusion

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