Research directions in processing and modeling Web-documents

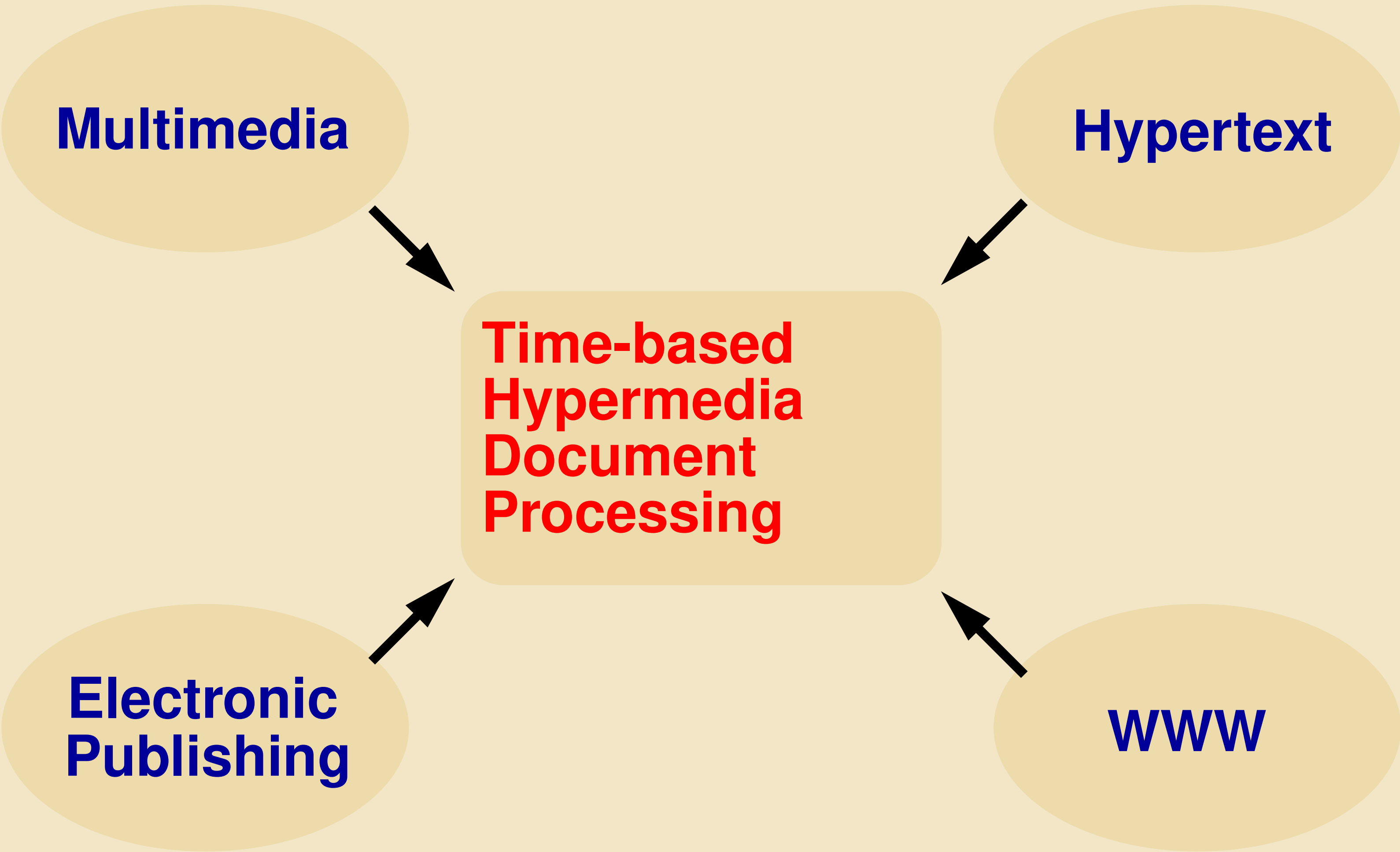
Jacco van Ossenbruggen

Multimedia and Human-Computer Interaction Group
CWI: Centrum voor Wiskunde en Informatica
Amsterdam, The Netherlands

Presentation Outline:

- Introduction
- Design dimensions
- Multiple delivery publishing
- Conclusion

Introduction



Design Dimensions

Applications versus documents

Content versus markup

Embedded versus external markup

Procedural versus declarative

Domain independent versus domain specific

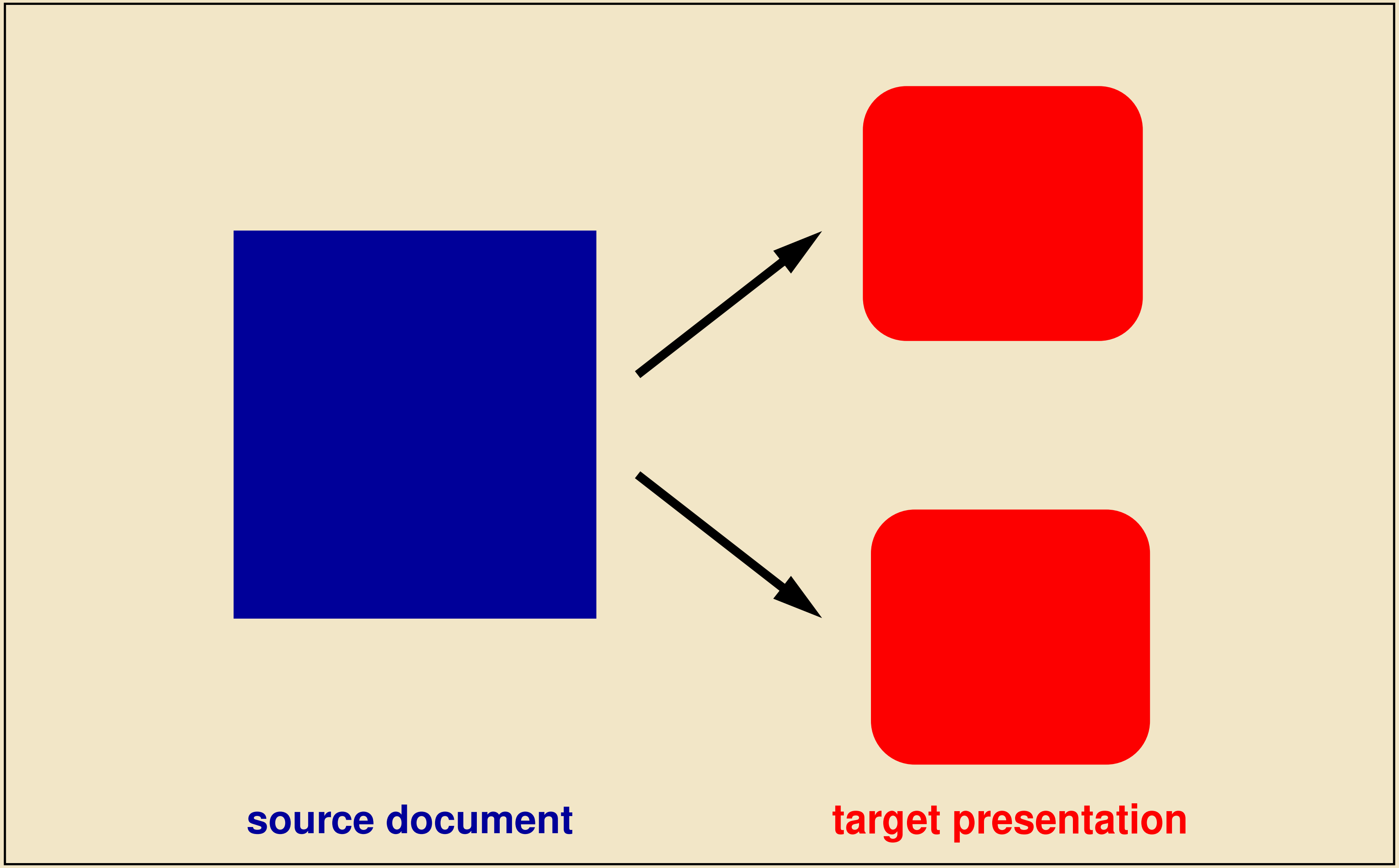
Layout-driven versus content-driven

Visual markup versus structured markup

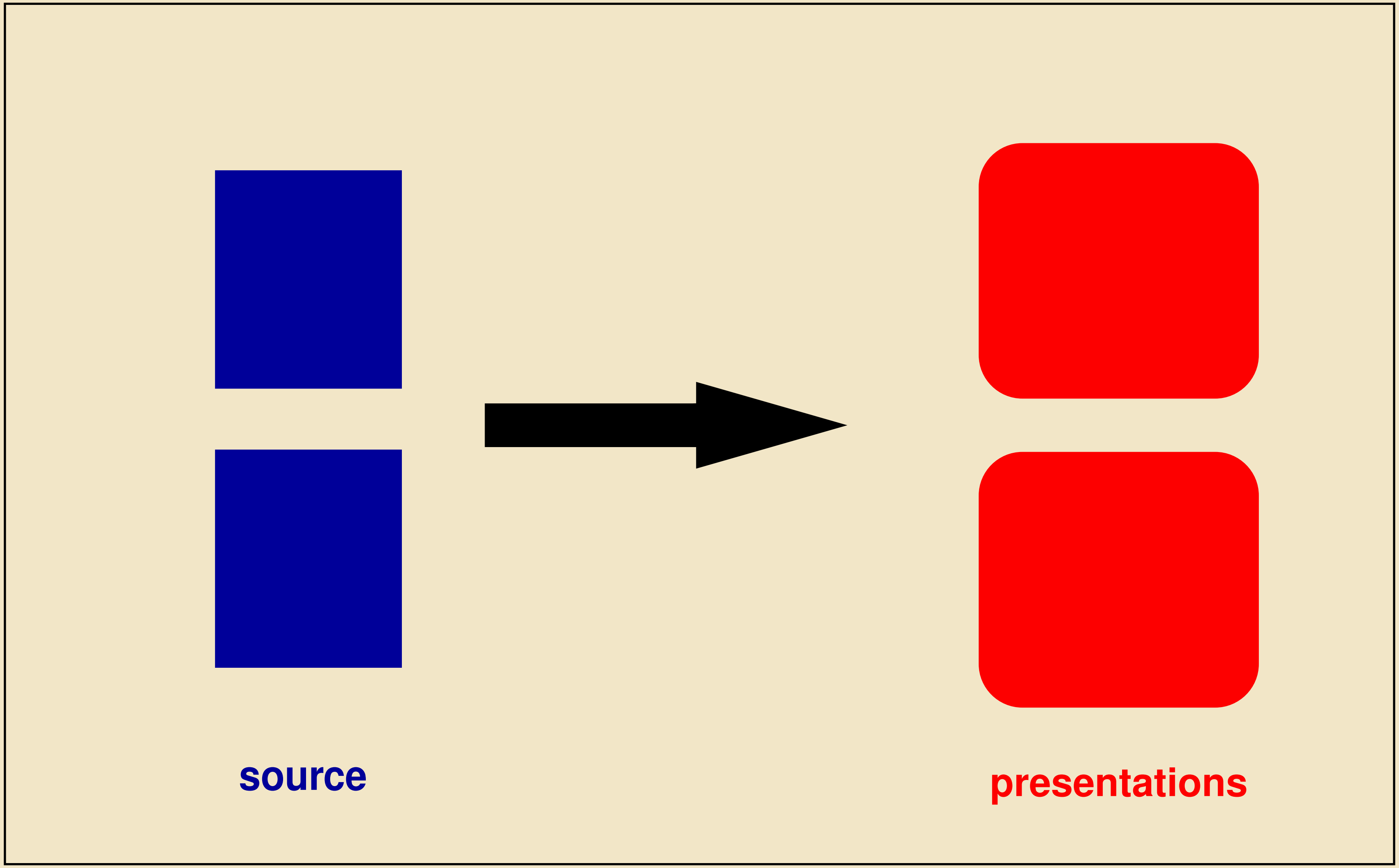
Document Classification

	Static	Dynamic
Linear	Text	Multimedia
Non-linear	Hypertext	Hypermedia

Multiple delivery publishing (1)



Multiple delivery publishing (2)



Multiple delivery publishing (3)

Source format:

- Structured, declarative markup
- Can be domain independent but...
- ...is usually tailored to a specific domain
- Provide sufficiently rich structure for style sheets and other processing

Presentation format:

- Visual, often procedural markup
- Can be platform/medium independent but...
- ... is usually tailored to a specific output medium/device
- Provide sufficient information to obtain high quality output

How do you classify your favorite document format?

Multiple delivery publishing (4)

Advantages:

- Longevity
- Reusability
- Flexibility & Tailorability

Disadvantages:

- Complexity
- More dependent on tools
- Training
- Initial investment

Multiple delivery publishing works best for content-driven material

Multiple delivery publishing on the Web

	W3C/HTML		
Markup	HTML		
Style	CSS		
Linking	<a href=		
Addressing	<a name=		
Extensions	ad hoc		

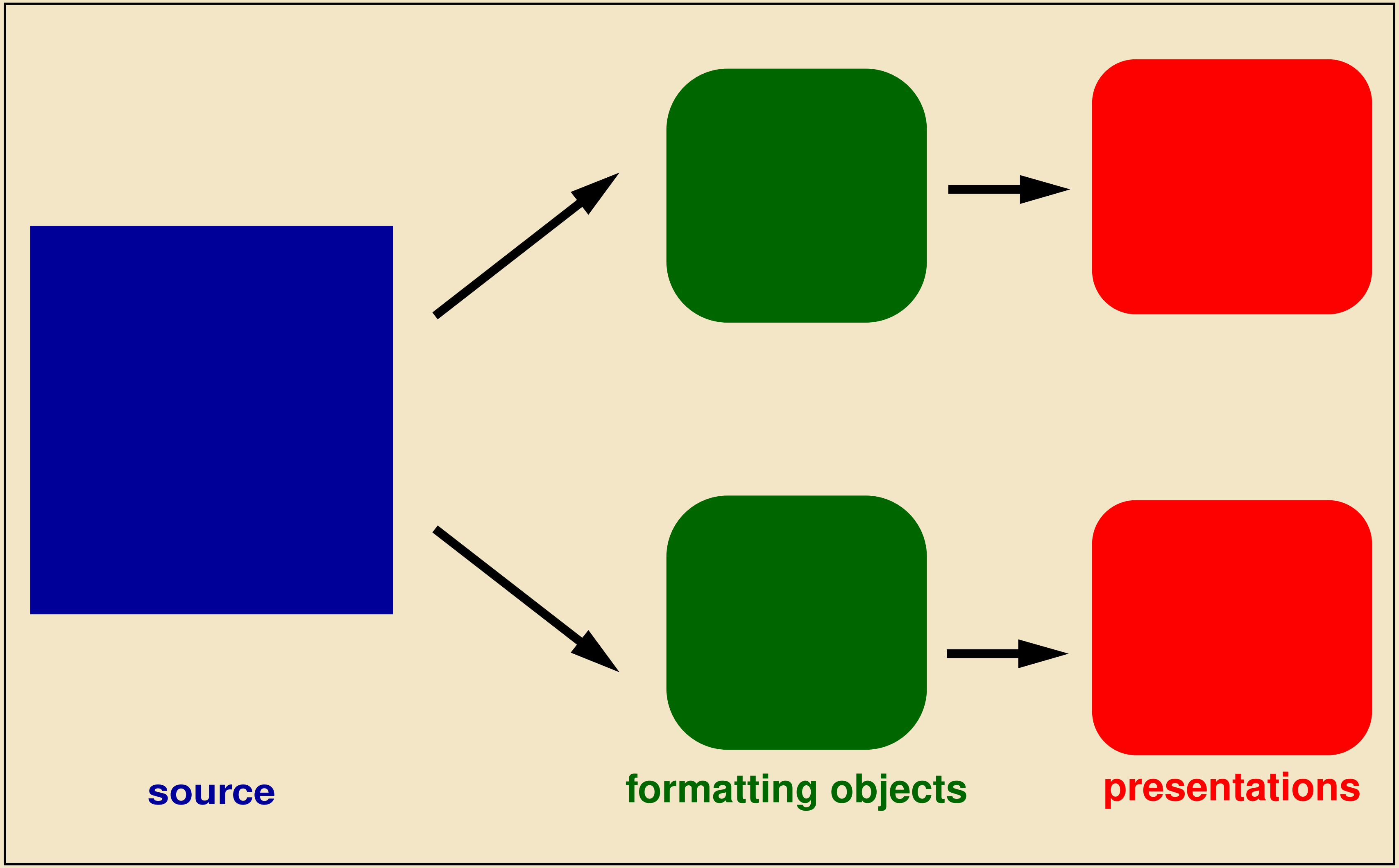
Multiple delivery publishing on the Web

	W3C/HTML		ISO/SGML
Markup	HTML		SGML
Style	CSS		DSSSL
Linking	<a href=		HyTime/TEI
Addressing	<a name=		HyTime/TEI
Extensions	ad hoc		Architectures

Multiple delivery publishing on the Web

	W3C/HTML	W3C/XML	ISO/SGML
Markup	HTML	XML	SGML
Style	CSS	(XSL)	DSSSL
Linking	<a href=	(XLink)	HyTime/TEI
Addressing	<a name=	(XPointer)	HyTime/TEI
Extensions	ad hoc	Namespaces	Architectures

Style sheets: Formatting objects



Style sheets: languages

CSS

- Selectors: tag, class, name/id,...
- Structure input tree == Structure output tree

DSSSL

- Selectors: Query language (SDQL)
- Transformation Language + Formatting Language

(XSL)

- Selectors: XSL patterns
- Transforms to formatting output vocabulary

Issues:

- Selectors overlap with XPointer:
  — “Does the style sheet query the document?” or... vice versa?
- Formatting versus transformation

Style sheets: Transformation (XTL?)

XML to XML

- Converting from in-house to standard DTD
- Up/down conversions
- Converting legacy (SGML) documents

XML to formatting object vocabulary

- HTML/CSS
- SMIL
- ...

XML to anything

- RTF
- ...

Hypertext

Visionaries: Bush, Engelbart, Nelson

Rich link structures: multi-headed, bidirectional, typed

Open Hypermedia Systems & non-embedded linking:

- no write access to create links
- do not require change to file formats
- extend existing applications

Halasz' seven issues

Dexter Hypertext Reference Model

Comparable to MDP: Adaptive hypertext

Multimedia

Layout

- Temporal (synchronization, scheduling)
- Spatial (not based on text-flow, scroll bars)

Authoring paradigms

- time lines
- flow charts
- scripts
- temporal constraints
- temporal document structure

Comparable to MDP: QoS

- document
- style/transformation sheet
- transport protocol

Hypermedia

multimedia hypertext

- example: HTML
- has links, but no time line

linear hypermedia

- example SMIL 1.0
- has links, but within single timeline

non-linear hypermedia

- example: Amsterdam Hypermedia Model
- has links among multiple time lines

Hypermedia and multiple delivery publishing

What is content, what is structure and what is style?

- Coarse-grained temporal alignment (`par/seq`)
- Fine-grained temporal alignment (`begin/end/dur`)
- Basic spatial layout (`regions`)
- Alternatives (`switch)`
- Transitions (not in SMIL)
- Background colors, anchor styles, etc (not in SMIL)

Which aspects of the document are likely to vary when different presentations are used?

When do you have sufficient information?

- authoring
- rendering
- presentation

Wrap up

Multiple delivery publishing

- Pro: longevity, reusability, flexibility
- Con: complexity, not suitable for layout-driven applications

Multiple delivery publishing on the Web

- XML is only the beginning
- We need style, transformation, linking, timing,...

What is the style of a hypermedia presentation?

- Time, space, links, alternative content
- Part of a hypermedia presentation is layout-driven