Plan for Today's Lecture(s)

- The Structural Perspective on Relationships
- Internal Structure in Documents
- Motivating "Document Engineering"
- The Document Engineering Method
- "Heavyweight" vs. "Lightweight" Modeling

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

8 October 2013
Lecture 12.1 – The Structural Perspective on Relationships

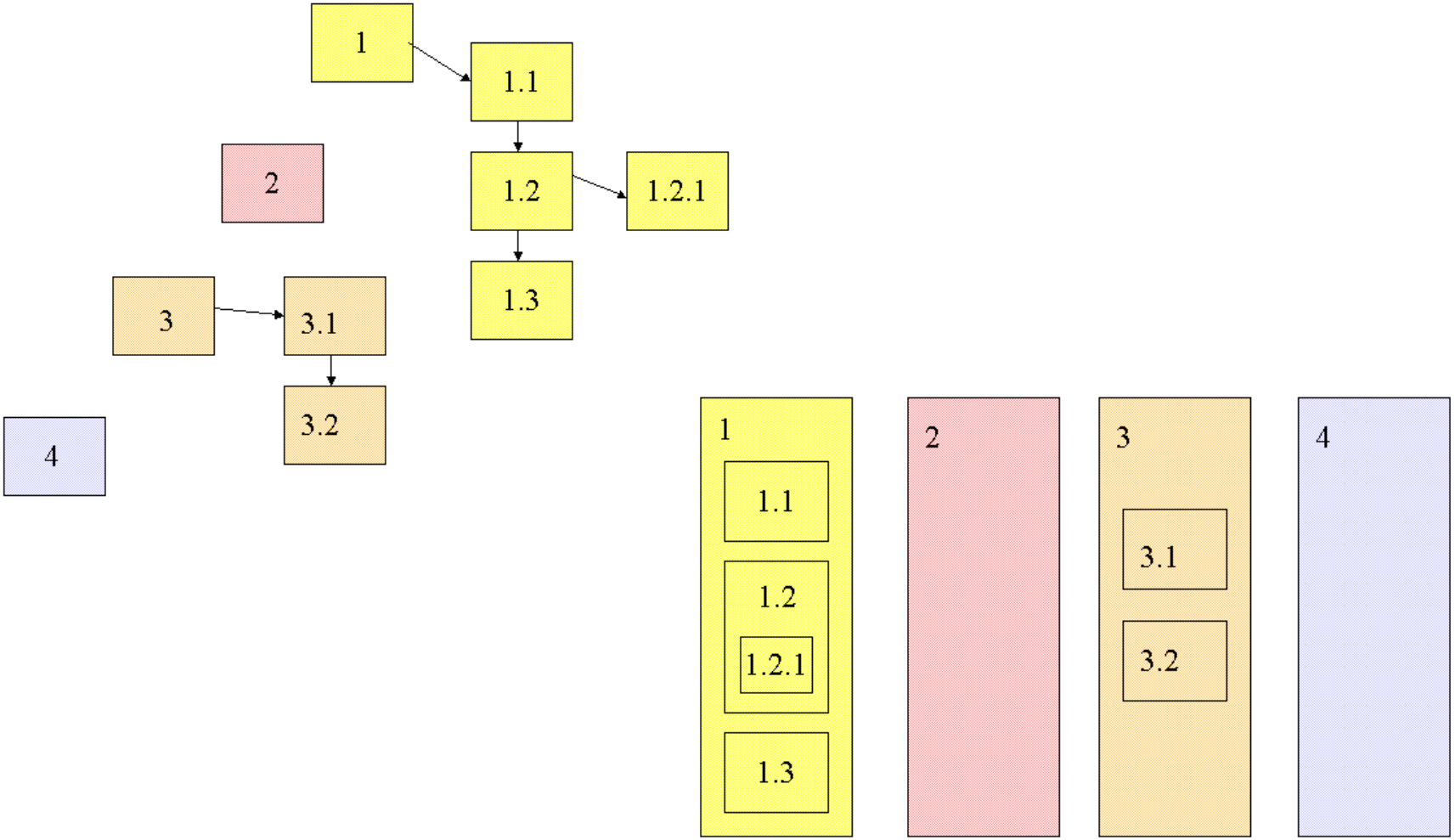
The Structural Perspective on Relationships

- Analyzing the association, arrangement, proximity, or connection between resources without primary concern for their meaning or the origin of these relationships
- Sometimes structure is all we know...and sometimes we ignore what we know about relationship semantics to focus on the generic aspect of structural connectivity

Internal and External Structure

- Resources can have INTERNAL structure as well as EXTERNAL structure that connects them to other resources
- We often make arbitrary decisions about how the granularity with which we describe the internal structure of a resource
- The boundaries we impose to identify resources determines whether some structure is internal or external with respect to them

Changing Resource Boundaries Changes External to Internal Structure



Intentional Structure

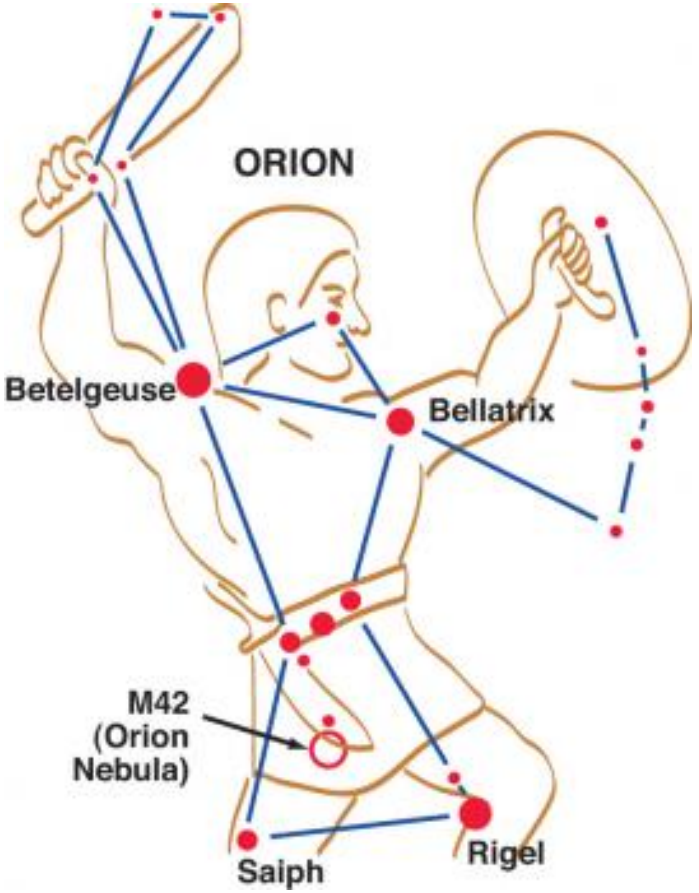
- In organizing systems we focus on INTENTIONAL structure created by people or computational agents programmed by people
- If structure isn't intentional, we can't reuse or recreate the organizing system with our own efforts

Structure, But Not Intentional

Making Implicit Structure Explicit

- Some organizing principles impose very little structure and the structure might be implicit
  - co-location / pile of resources used together
  - arrangement by frequency of use
- Coordinate systems enable absolute description of location, or we can use relative descriptions
  - in, on, above, below, in front of, behind…
- Statistical techniques can make explicit the structure implied by frequency distribution or correlations between properties of resources

Sometimes Implicit Structure is Apparent… but not Organized



Orion's Belt: Alnitak (736 LY), Alnilam (1340LY), Mintaka (915 LY)

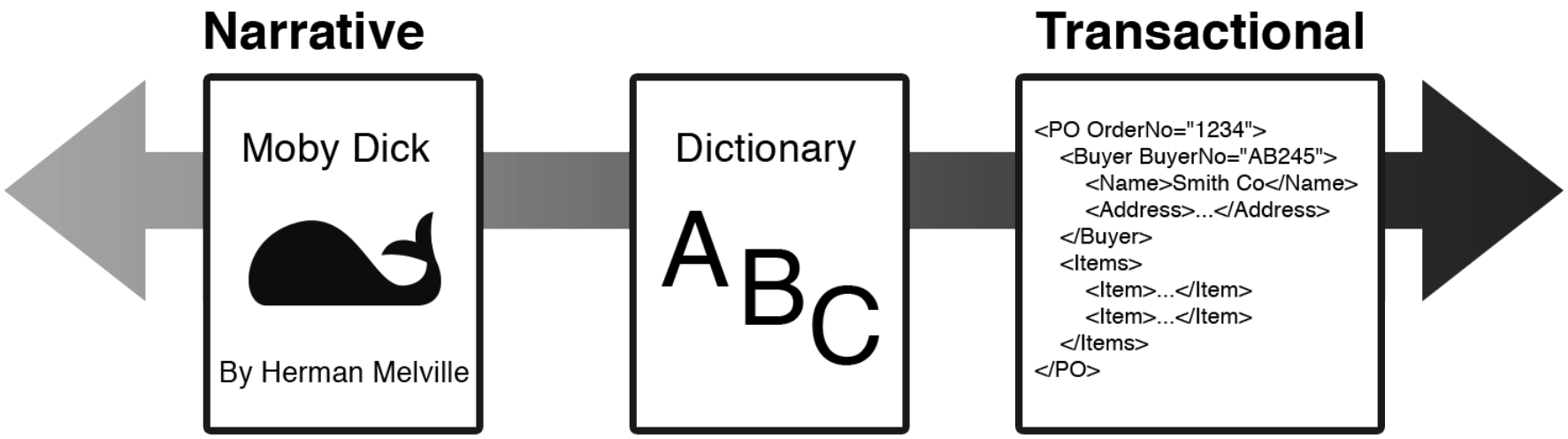
UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

8 October 2013
Lecture 12.2 – Internal Structure in Documents

The Document Type Spectrum



Structure in Narrative Document Types

- NARRATIVE documents typically have relatively little internal structure, and that usually varies across instances of the same document type

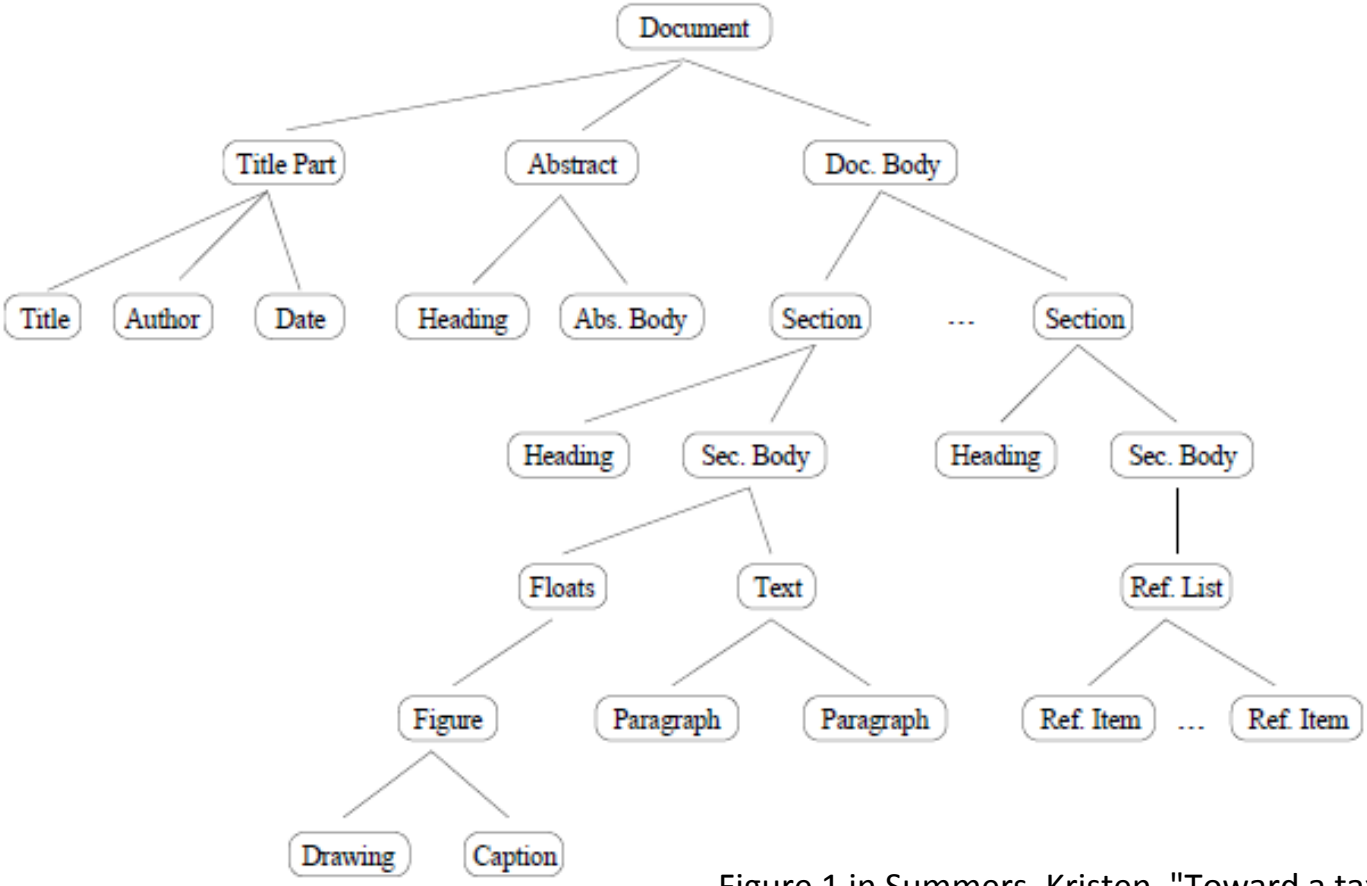


Figure 1 in Summers, Kristen. "Toward a taxonomy of logical document structures (1995)

The Exception that Proves the Rule

- There are some notable exceptions for specialized narrative document types like "theatrical play" where there are highly conventional structural divisions and associated presentation rules
- Or should we say that the specialized use cases or context of use for plays requires us to impose more structure than most narrative document types need

The Exception that Proves the Rule

```
<play>
  <author>Shakespeare</author> <title>Macbeth</
  title> <act number="I“>       <scene
  number="VII">
  <title>Macbeth's castle</title>
   <verse>Will I with wine and wassail...</verse>…
  </scene>…
   </act>…
  </play>
```

Structure in Hybrid Document Types

- HYBRID document types are often called "semi-structured“ because they have more structure than narrative ones but less than fully-structured transactional ones
- The extent of explicit structure (“mixed content”) in semi-structured document types is an important design decision

Encyclopedia:
Semi-Structured Document Type

History [edit]

Early history [edit]

The site of today's City of Berkeley was the territory of the Chochenyo/Huchiun band of the Ohlone people when the first Europeans arrived.[3] Evidence of their existence in the area include pits in rock formations, which they used to grind acorns, and a shellmound, now mostly leveled and covered up, along the shoreline of San Francisco Bay at the mouth of Strawberry Creek. Other artifacts were discovered in the 1950s in the downtown area during remodeling of a commercial building, near the upper course of the creek.

The first people of European descent (most of whom were born in America, and many of whom were of mixed ancestry[citation needed]) arrived with the De Anza Expedition in 1776. Today, this is noted by signage on Interstate 80, which runs along the San Francisco Bay shoreline of Berkeley. The De Anza Expedition led to establishment of the Spanish Presidio of San Francisco at the entrance to San Francisco Bay (the *Golden Gate*), which is due west of Berkeley. Luís Peralta was among the soldiers at the Presidio. For his services to the King of Spain, he was granted a vast stretch of land on the east shore of San Francisco Bay (the *contra costa*, "opposite shore") for a ranch, including that portion that now comprises the City of Berkeley.

Luís Peralta named his holding "Rancho San Antonio." The primary activity of the ranch was raising cattle for meat and hides, but hunting and farming were also pursued. Eventually, Peralta gave portions of the ranch to each of his four sons. What is now Berkeley lies mostly in the portion that went to Peralta's son Domingo, with a little in the portion that went to another son, Vicente. No artifact survives of the ranches of Domingo or Vicente, although their names have been preserved in the naming of Berkeley streets (Vicente, Domingo, and Peralta). However, legal title to all land in the City of Berkeley remains based on the original Peralta land grant.

The Peraltas' Rancho San Antonio continued after Alta California passed from Spanish to Mexican sovereignty after the Mexican War of Independence. However, the advent of U.S. sovereignty after the Mexican–American War, and especially, the Gold Rush, saw the Peralta's lands quickly encroached on by squatters and diminished by dubious legal proceedings. The lands of the brothers Domingo and Vicente were quickly reduced to reservations close to their respective ranch homes. The rest of the land was surveyed and parceled out to various American claimants (See Kellersberger's Map).

Politically, the area that became Berkeley was initially part of a vast Contra Costa County. On March 25, 1853, Alameda County was created by division of Contra Costa County, as well as from a small portion of Santa Clara County.

The City of Berkeley highlighted within Alameda County

Coordinates: 37°52′18″N 122°16′22″W

Country	United States
State	California
County	Alameda
Incorporated	April 4, 1878
Government	
• **Type**	Mayor and City Council
• **Mayor**	Tom Bates
• **Councilmembers**	District 1: Linda Maio District 2: Darryl Moore District 3: Maxwell Anderson District 4: Jesse Arreguin District 5: Laurie Capitelli District 6: Susan Wengraf District 7: Kriss Worthington District 8: Gordon Wozniak
• **State Senate**	Loni Hancock (D) District 9
• **State Assembly**	Nancy Skinner (D) District 14
• **U. S. Congress**	Barbara Lee (D) District 9
Area[1]	
• **Total**	17.596 sq mi (45.833 km^2)
• **Land**	10.470 sq mi (27.118 km^2)
• **Water**	7.226 sq mi (18.716 km^2) 40.83%
Elevation	0–1,320 ft (0–400 m)

Structure in Transactional Documents

- TRANSACTIONAL documents are completely and regularly structured because they contain many specific information components that are distinguished by their content type
- These components are often organized in aggregate structures that can occur many times in the same document because they are typically created by repeated automated processes
- For example, an invoice might contain a long list of purchases, each of which identifies a buyer, an item, a quantity, and a total price

Call Center Log
Transactional Document Type

vru-line	call_id	customer_id	priority	type	date	vru_entry	vru_exit	vru_time	q_start	q_exit	q_time	outcome	ser_start	ser_exit	ser_time	server
AA0101	44749	27644400	2	PS	990901	11:45:33	11:45:39	6	11:45:39	11:46:58	79	AGENT	11:46:57	11:51:00	243	DORIT
AA0101	44750	12887816	1	PS	990905	14:49:00	14:49:06	6	14:49:06	14:53:00	234	AGENT	14:52:59	14:54:29	90	ROTH
AA0101	44967	58660291	2	PS	990905	14:58:42	14:58:48	6	14:58:48	15:02:31	223	AGENT	15:02:31	15:04:10	99	ROTH
AA0101	44968	0	0	NW	990905	15:10:17	15:10:26	9	15:10:26	15:13:19	173	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44969	63193346	2	PS	990905	15:22:07	15:22:13	6	15:22:13	15:23:21	68	AGENT	15:23:20	15:25:25	125	STEREN
AA0101	44970	0	0	NW	990905	15:31:33	15:31:47	14	00:00:00	00:00:00	0	AGENT	15:31:45	15:34:16	151	STEREN
AA0101	44971	41630443	2	PS	990905	15:37:29	15:37:34	5	15:37:34	15:38:20	46	AGENT	15:38:18	15:40:56	158	TOVA
AA0101	44972	64185333	2	PS	990905	15:44:32	15:44:37	5	15:44:37	15:47:57	200	AGENT	15:47:56	15:49:02	66	TOVA
AA0101	44973	3.06E+08	1	PS	990905	15:53:05	15:53:11	6	15:53:11	15:56:39	208	AGENT	15:56:38	15:56:47	9	MORIAH
AA0101	44974	74780917	2	NE	990905	15:59:34	15:59:40	6	15:59:40	16:02:33	173	AGENT	16:02:33	16:26:04	1411	ELI
AA0101	44975	55920755	2	PS	990905	16:07:46	16:07:51	5	16:07:51	16:08:01	10	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44976	0	0	NW	990905	16:11:38	16:11:48	10	16:11:48	16:11:50	2	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44977	33689787	2	PS	990905	16:14:27	16:14:33	6	16:14:33	16:14:54	21	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44978	23817067	2	PS	990905	16:19:11	16:19:17	6	16:19:17	16:19:39	22	AGENT	16:19:38	16:21:57	139	TOVA
AA0101	44764	0	0	PS	990901	15:03:26	15:03:36	10	00:00:00	00:00:00	0	AGENT	15:03:35	15:06:36	181	ZOHARI
AA0101	44765	25219700	2	PS	990901	15:14:46	15:14:51	5	15:14:51	15:15:10	19	AGENT	15:15:09	15:17:00	111	SHARON
AA0101	44766	0	0	PS	990901	15:25:48	15:26:00	12	00:00:00	00:00:00	0	AGENT	15:25:59	15:28:15	136	ANAT
AA0101	44767	58859752	2	PS	990901	15:34:57	15:35:03	6	15:35:03	15:35:14	11	AGENT	15:35:13	15:35:15	2	MORIAH
AA0101	44768	0	0	PS	990901	15:46:30	15:46:39	9	00:00:00	00:00:00	0	AGENT	15:46:38	15:51:51	313	ANAT

The Structure of a Web Page

- Conceptually convenient to think of a web page as a unitary object with a unique identifier
- But individual web pages are highly structured
- Static internal structure is encoded in the markup of the web page
  - As HTML
  - As XHTML
  - As XHTML, with additional XML vocabularies identified by namespace

The Structure of a Web Page

- Scripts can be used to generate or modify content when the page is loaded or when a user interacts with it
- Dynamic web pages might have a static structure into which content is "poured" or the structure might also be generated dynamically
- There may be hyperlinks within a web page, like a table of contents to facilitate navigation within a long page that would otherwise require lots of scrolling

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

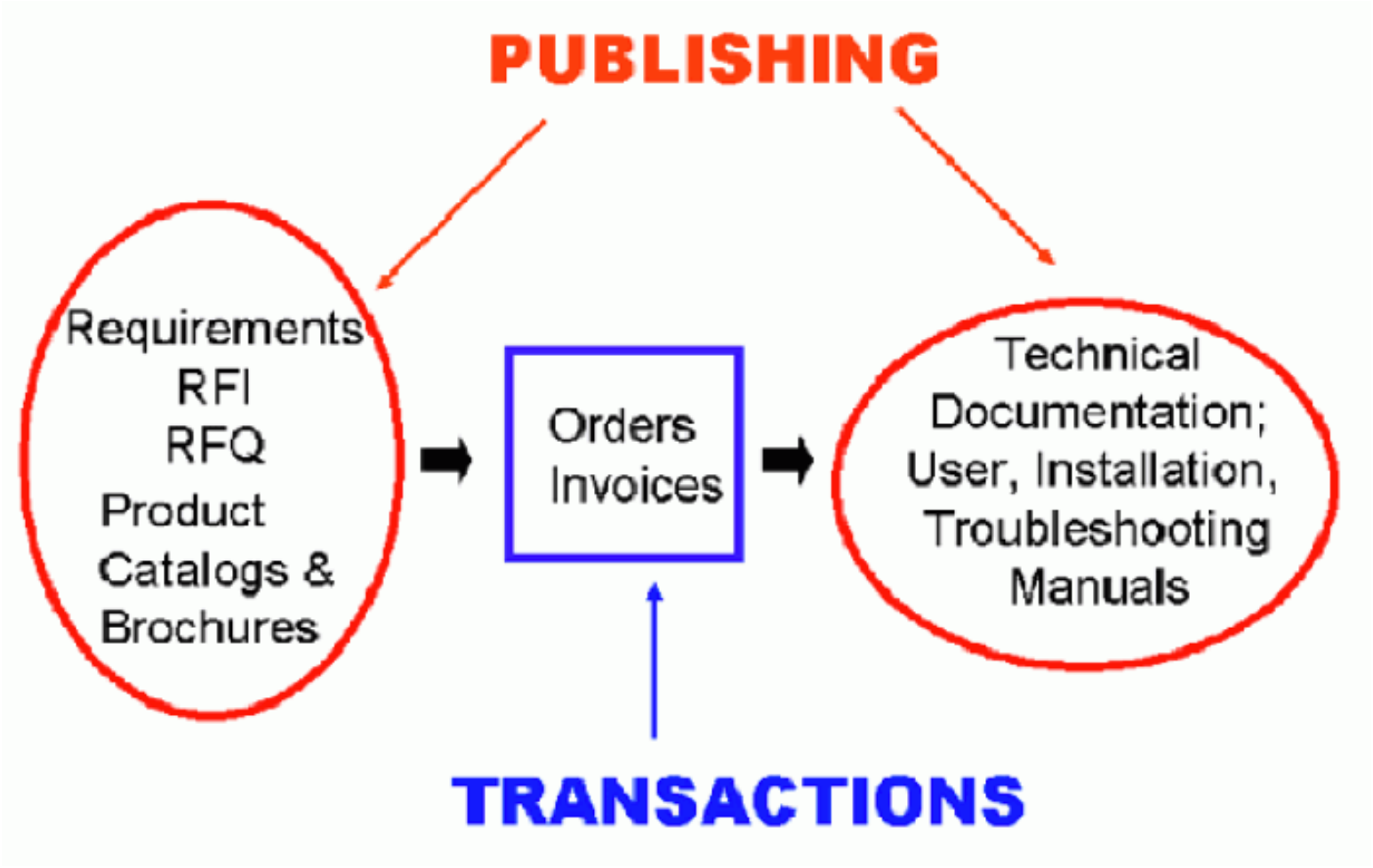
Robert J. Glushko
glushko@berkeley.edu
@rjglushko

8 October 2013
Lecture 12.3 – Motivating
"Document Engineering"

Documents vs. Data

- Many people have contrasted "documents" and "data" and concluded that documents and data cannot be understood and handled with the same terminology, techniques, and tools.
- This document vs. data distinction is embedded and reinforced in textbooks, technology, and product marketing

Mixing Data and Documents



Catalog:
Database That Contains Documents

Industrial Or Light Weight Bags On A Roll

Bags are perforated allowing easy tear off for in-store or assembly line use. Choose industrial 2-mil or extra heavy 4-mil for parts fittings and hardware. Lightweight bags are .5 mil and ideal for produce and lighter weight items. Table or wall mount dispenser available below.

Industrial Bags On A Roll

Size (In.)	Bags Per Roll	Part No. 2 Mil	Price 2 Mil	Part No. 4 Mil	Price 4 Mil
4 x 6	1000	88400LU	25.55	88430LU	45.55
6 x 9	1000	88403LU	39.15	88433LU	59.90
8 x 10	1000	88406LU	47.15	88436LU	85.65
10 x 12	1000	88409LU	68.25	88439LU	125.30

Discount per part no.: Less 5% 12-23 rolls; 15% 24 rolls or more.

Lightweight Bags On A Roll

Size (In.)	Bags Per Roll	Part No. .5 Mil	Price Per Carton of 2 Rolls 1-11	12-23	24+
10 x 15	2000	88080LU	52.80	50.16	44.88
10 x 20	1500	88085LU	52.80	50.16	44.88
Dispenser		88090LU		17.80 each	

Lay-Flat Poly Tubing Rolls

Simply cut tubing to your exact length and seal with the Consolidated Impulse Heat Sealer found on page 80. Choose 2 mil or 4 mil tubing stock. Ideal for a variety of different size parts. FDA approved.

2 Mil

Part No.	W x L (In. x Ft.)	Price/Roll	Part No.	W x L (In. x Ft.)	Price/Roll
89965LU	2 x 2100	33.66	89972LU	12 x 2100	101.18
89967LU	3 x 2100	39.40	89973LU	14 x 2100	119.86
89968LU	4 x 2100	55.90	89974LU	16 x 2100	138.60
89969LU	5 x 2100	63.85	89975LU	18 x 2100	147.72
89971LU	6 x 2100	62.66	89976LU	20 x 2100	160.05
89979LU	8 x 2100	78.64	89940LU	24 x 1700	168.43
89981LU	10 x 2100	94.40	89941LU	36 x 1100	163.49

4 Mil

Part No.	W x L (In. x Ft.)	Price/Roll	Part No.	W x L (In. x Ft.)	Price/Roll
89980LU	2 x 1050	33.66	89993LU	12 x 1050	101.18
89982LU	3 x 1050	39.40	89994LU	14 x 1050	119.86
89983LU	4 x 1050	55.90	89996LU	16 x 1050	138.60
89984LU	5 x 1050	63.85	89997LU	18 x 1050	147.72
89986LU	6 x 1050	62.66	89998LU	20 x 1050	160.05
89988LU	8 x 1050	78.64	89942LU	24 x 850	168.43
89992LU	10 x 1050	94.40	89943LU	36 x 550	163.49

Discount per part no.: Less 5% 5-11 rolls; 10% 12-23 rolls; 15% 24 rolls or more.

Encyclopedia:
Document that Contains Databases

History [edit]

Early history [edit]

The site of today's City of Berkeley was the territory of the Chochenyo/Huchiun band of the Ohlone people when the first Europeans arrived.[3] Evidence of their existence in the area include pits in rock formations, which they used to grind acorns, and a shellmound, now mostly leveled and covered up, along the shoreline of San Francisco Bay at the mouth of Strawberry Creek. Other artifacts were discovered in the 1950s in the downtown area during remodeling of a commercial building, near the upper course of the creek.

The first people of European descent (most of whom were born in America, and many of whom were of mixed ancestry[citation needed]) arrived with the De Anza Expedition in 1776. Today, this is noted by signage on Interstate 80, which runs along the San Francisco Bay shoreline of Berkeley. The De Anza Expedition led to establishment of the Spanish Presidio of San Francisco at the entrance to San Francisco Bay (the *Golden Gate*), which is due west of Berkeley. Luís Peralta was among the soldiers at the Presidio. For his services to the King of Spain, he was granted a vast stretch of land on the east shore of San Francisco Bay (the *contra costa*, "opposite shore") for a ranch, including that portion that now comprises the City of Berkeley.

Luís Peralta named his holding "Rancho San Antonio." The primary activity of the ranch was raising cattle for meat and hides, but hunting and farming were also pursued. Eventually, Peralta gave portions of the ranch to each of his four sons. What is now Berkeley lies mostly in the portion that went to Peralta's son Domingo, with a little in the portion that went to another son, Vicente. No artifact survives of the ranches of Domingo or Vicente, although their names have been preserved in the naming of Berkeley streets (Vicente, Domingo, and Peralta). However, legal title to all land in the City of Berkeley remains based on the original Peralta land grant.

The Peraltas' Rancho San Antonio continued after Alta California passed from Spanish to Mexican sovereignty after the Mexican War of Independence. However, the advent of U.S. sovereignty after the Mexican–American War, and especially, the Gold Rush, saw the Peralta's lands quickly encroached on by squatters and diminished by dubious legal proceedings. The lands of the brothers Domingo and Vicente were quickly reduced to reservations close to their respective ranch homes. The rest of the land was surveyed and parceled out to various American claimants (See Kellersberger's Map).

Politically, the area that became Berkeley was initially part of a vast Contra Costa County. On March 25, 1853, Alameda County was created by division of Contra Costa County, as well as from a small portion of Santa Clara County.

The City of Berkeley highlighted within Alameda County

Coordinates: 37°52′18″N 122°16′22″W

Country	United States
State	California
County	Alameda
Incorporated	April 4, 1878
Government	
• **Type**	Mayor and City Council
• **Mayor**	Tom Bates
• **Councilmembers**	District 1: Linda Maio District 2: Darryl Moore District 3: Maxwell Anderson District 4: Jesse Arreguin District 5: Laurie Capitelli District 6: Susan Wengraf District 7: Kriss Worthington District 8: Gordon Wozniak
• **State Senate**	Loni Hancock (D) District 9
• **State Assembly**	Nancy Skinner (D) District 14
• **U. S. Congress**	Barbara Lee (D) District 9
Area[1]	
• **Total**	17.596 sq mi (45.833 km^2)
• **Land**	10.470 sq mi (27.118 km^2)
• **Water**	7.226 sq mi (18.716 km^2) 40.83%
Elevation	0–1,320 ft (0–400 m)

Document Analysis

- Documents are Artifacts or Renditions that combine content, structure and appearance
- The goal of document analysis is a model of a document's content and structure that is separate from its presentational characteristics
- The optimal prescriptive schema for a set of documents is one that best satisfies the requirements of current and prospective users for carrying out specific tasks with new instances
- Finally, one or more stylesheets can be used to assign formatting or rendering characteristics in a consistent manner to any valid document

"Document Analysis" Methodology

- Heritage: publishing, literary analysis, graphical design
- Scope: One document type at a time
- Reuse focus: identify "boilerplate" content and repeating structural elements
- Heuristic rather than formal techniques
- Canonical textbook: Maler and Andaloussi. *Developing SGML DTDs: From Text to Model to Markup* (1996)

Data-Centric Analysis

- Goal is to understand and describe the properties and relationships between information components or objects
- This understanding is represented in conceptual models that organize the components efficiently to support a broad range of contexts or applications
- The conceptual model is also typically called a schema, but this is generally meant to be a "database schema" rather than a "document schema"

"Data Modeling" Methodology

- Heritage: philosophy, linguistics, systems analysis
- Scope: multiple interrelated document types
- Reuse focus: identify the overlapping content in transformationally- or derivationally-related document types
- Prescriptive and formal approach (e.g., schema normalization)
- Canonical textbook: C. J. Date, *An Introduction to Database Systems* (8 editions since 1980).

The Document Type Spectrum: A Continuum

- There is systematic and continuous variation in document instances and types and there is no clear boundary between “documents” and “data”
- But the traditional tools, terminology, and techniques for analyzing documents and data have made it into a chasm
- How do we cross the chasm?

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

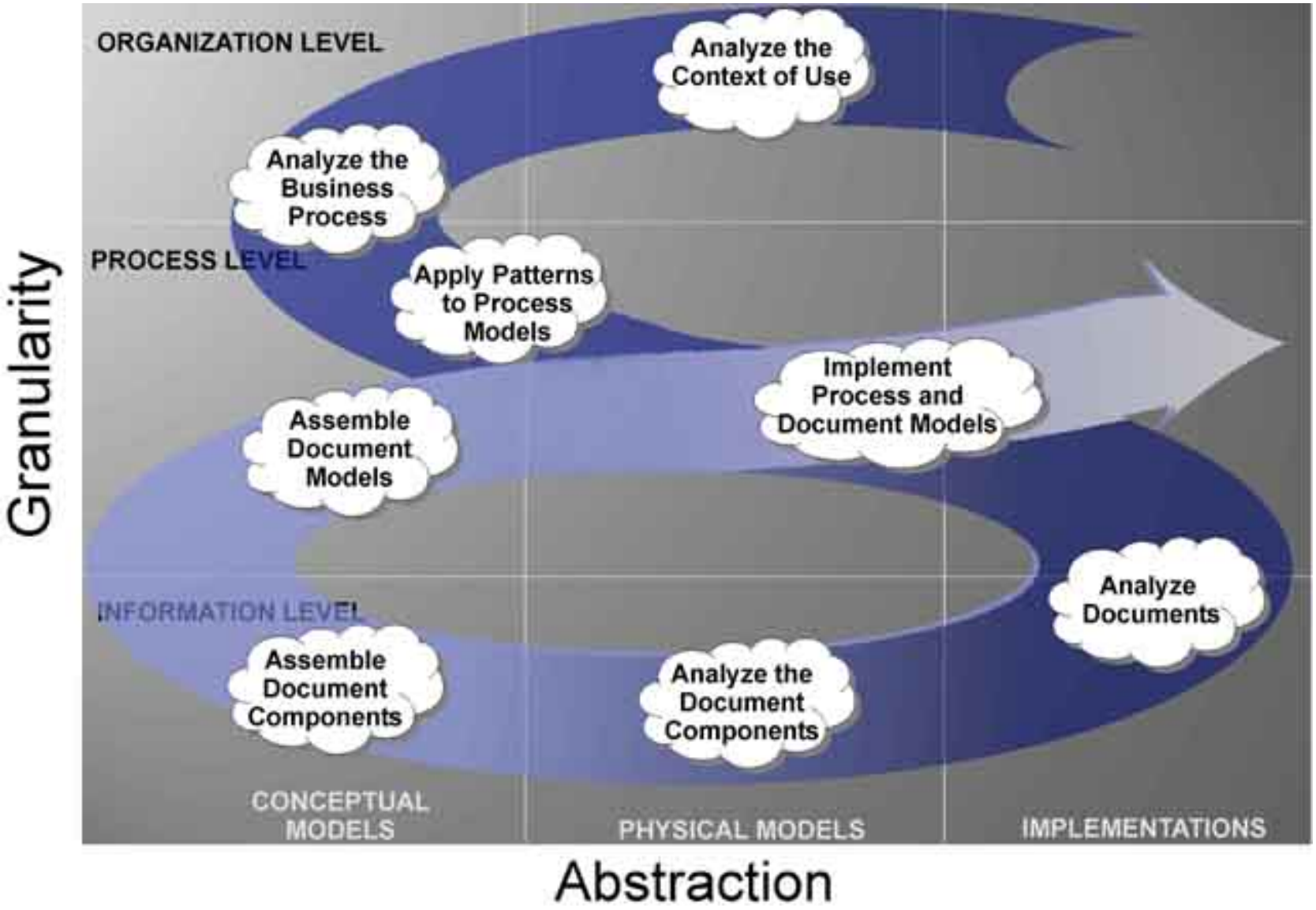
Robert J. Glushko
glushko@berkeley.edu
@rjglushko

8 October 2013
Lecture 12.4 –
The "Document Engineering" Method

Crossing the Chasm with "Document Engineering"

- The emerging discipline of Document Engineering unifies the document analysis and data modeling disciplines
- Document Engineering supports the specification, design, and implementation of the narrative and transactional documents needed in document-centric applications
- It is a synthesis of information and systems analysis, business process modeling, content management, and distributed computing
- Document Engineering requires and reinforces patterns of information description and exchange from the perspective of business models, business processes, and information systems and services design

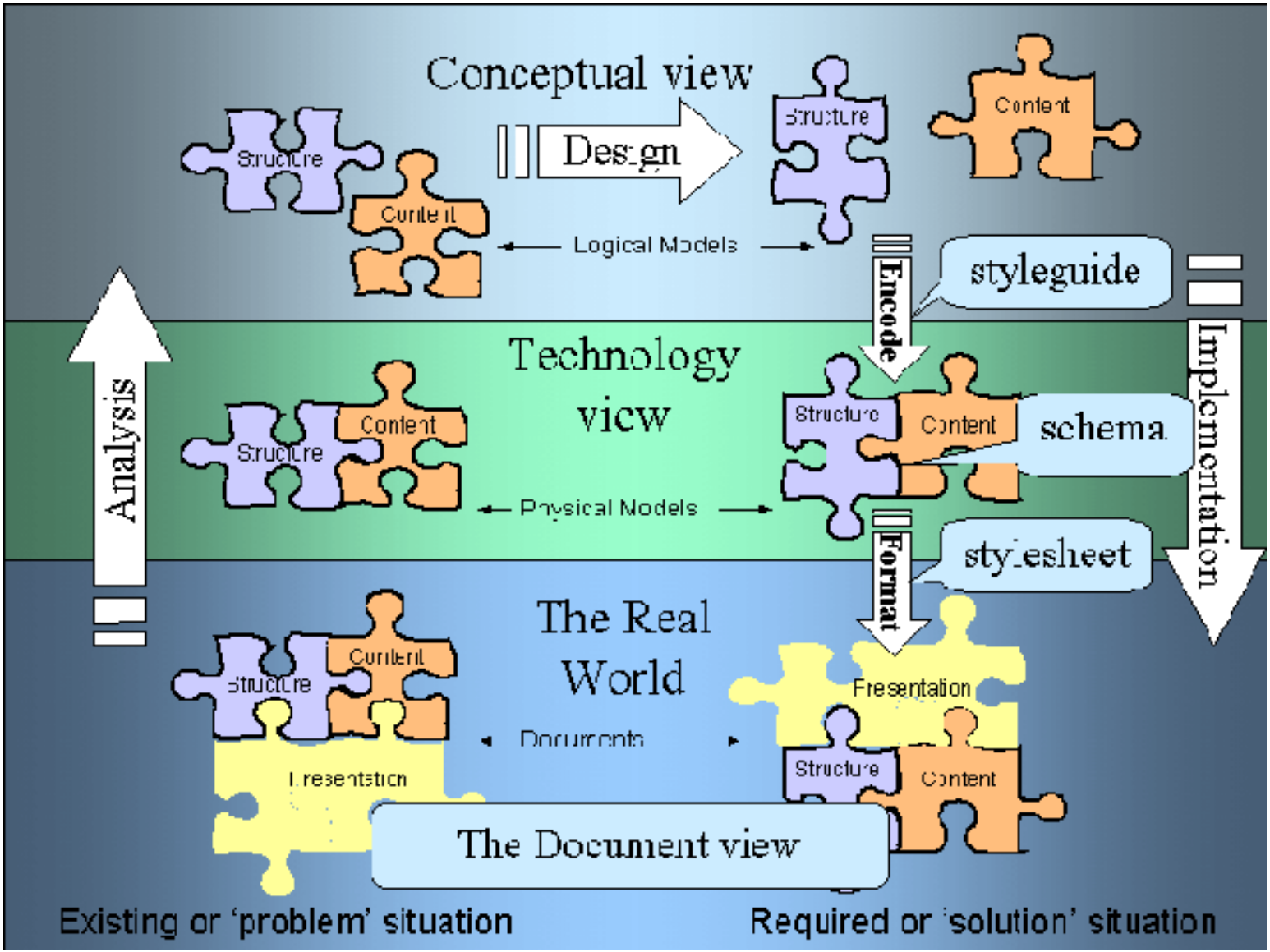
The Document Engineering Approach



Document Engineering's Key Ideas

- Document Engineering harmonizes the terminology of document analysis and data modeling and emphasizes what they have in common rather than highlighting their differences:
  - Identifying the presentational, content, and structural components
  - Eliminating synonymy and homonymy
  - Identifying and organizing the "good" content components
  - Assembling hierarchical document models to organize components to meet requirements for a specific context

The Document Engineering Approach



From Physical to Conceptual Models (1)

- When we analyze information sources: interviews, documents, sets of data whatever - our goal is to identify and describe the "significant things" or the "information components" and their characteristics or attributes
- If you conduct an interview your source will naturally talk about information components as abstract pieces of information
- But when you analyze documents the information components aren't as immediately apparent because they are contained in structures and rendered in some presentation

From Physical to Conceptual Models (2)

- So we have to remove the presentational information and dis-assemble the structural information to find the content information that is our highest priority
- As we take away presentation and structure, we are abstracting away or generalizing from a physical implementation and creating our first conceptual or logical model of the information components

Finding the Right Documents to Analyze (1)

- Identifying all the potentially relevant documents or information sources is inherently an iterative task
- We locate documents, which may refer or link to other documents, and we locate people who work with the documents, who may refer or link to other documents or people
- If there are lots of documents of a particular type, we have to be concerned about representiveness and selection biases

Finding the Right Documents to Analyze (2)

- Avoid the temptation to assume that job titles and formal organizational structure reflect what people actually do
- Likewise, avoid the temptation to assume that the names given to documents fit the people, tasks, and organizations in which we locate them
- Regardless of its title, make sure a document is being used before you conclude it is important

Extracting Presentation Rules

- Presentation affects structure and content by applying transformation rules to them
- To understand the structure and content we must identify and record what the rules of the transformation were
- Explicit transform rules can be encoded in templates, stylesheets or source code

Sometimes Rules Can't be Extracted

- No access to source formats or source code
- Rules may be inaccessible in source formats ("override" formatting in word processors instead of style tags)
- Rules don't exist or are inconsistently followed (author has "fontitis" with "ransom note" presentation style)

“Ransom Note” or “Fontitis” Style

Using Correlations or Conventions

- Color, pitch, other perceptual dimensions can be correlated with semantic distinctions
- Type size is correlated with structural hierarchy
- Content types can have characteristic layouts or text attributes
- Adjacency can suggest a semantic relationship, like that between figure and caption
- Presentation order is sometimes semantically significant

Presentation that Conveys Semantics

Abbreviate (ăbrī·vi,ēt), *v.*, also 5–7 **abreviate**. [f. ABBREVIATE *ppl. a.*; or on the analogy of vbs. so formed; see -ATE. A direct representative of L. *abbreviāre*; as ABRIDGE, and the obs. ABREVY, represent it indirectly, through OFr. *abregier* and mid. Fr. *abréviер*. Like the latter, *abbreviate*, was often spelt *a-breviate* in 5–7.] To make shorter, shorten, cut short in any way.

1530 PALSGR., I abrevyate: I make a thynge shorte, *Je abrege*. 1625 BACON *Essays* xxiv. 99 (1862) But it is one Thing to Abbreviate by Contracting, Another by Cutting off.

†1. *trans.* To make a discourse shorter by omitting details and preserving the substance; to abridge, condense. *Obs.*

a 1450 *Chester Pl.* I. 2 (Sh. Soc.) This matter he abbrevited into playes twenty-foure. 1592 GREENE *Conny catching* III. 16 The queane abreuiated her discourse. 1637 RALEIGH *Mahomet* 34 Abreviated out of two Arabique writers translated into Spanish. 1672 MANLEY *Interpreter* pref., I have omitted several Matters . . contracted and abbreviated Others.

†b. To make an abstract or brief of, to epitomize. *Obs.*

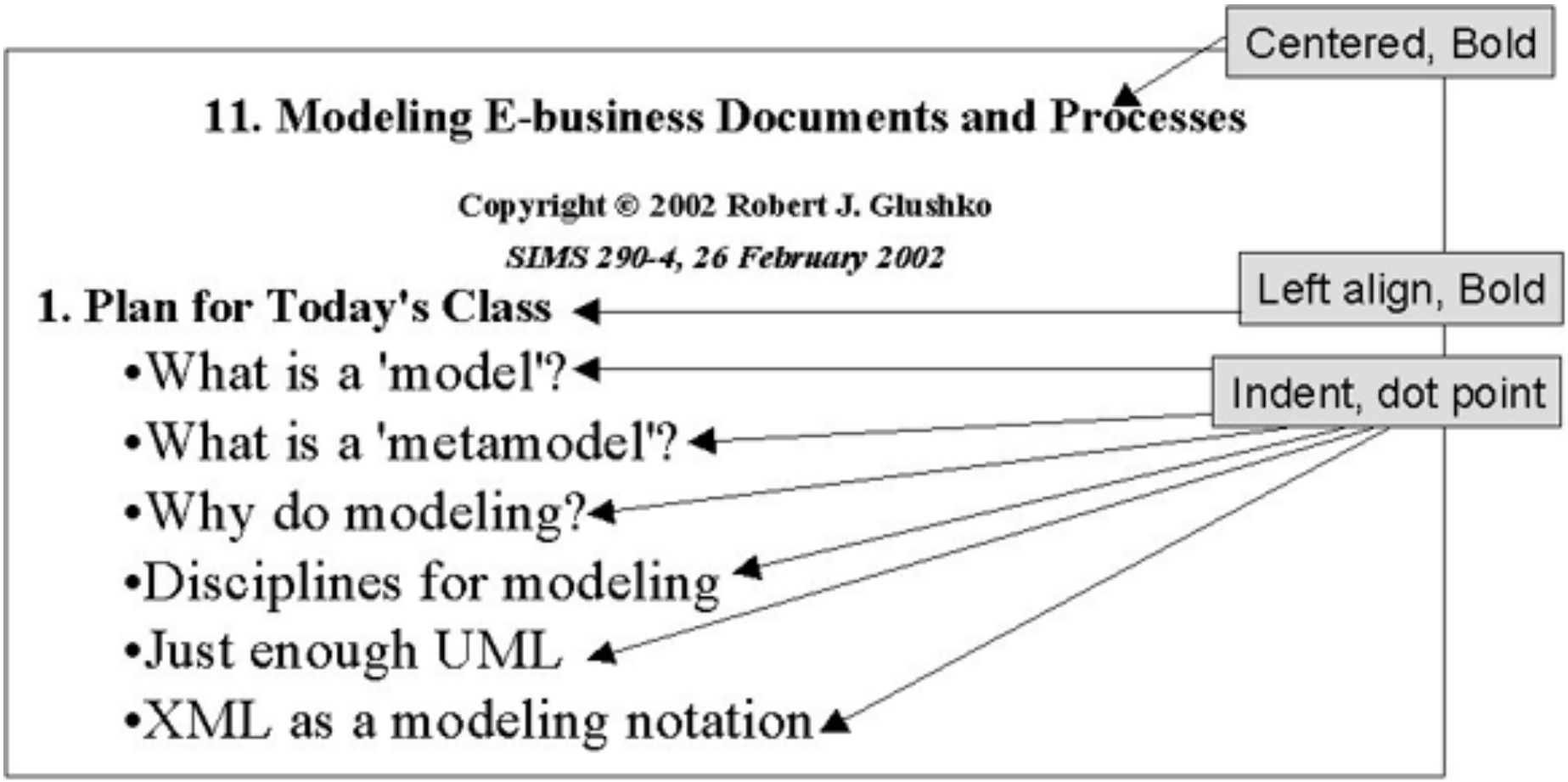
c 1450 TREVISA *Higden's Polychr.* I. 21 (Rolls Ser.) Trogus Pompeius, in hys xl^{ti} iiij. bookes, allemoste of alle the storyes

Presentation that Hides Content Components

Address:
Line 1: _____
Line 2: _____
City: _____ State: _____ ZipCode: _____

- “Line 1" and “Line 2" are presentation labels that are not useful for any purpose other than printing out an address label
- They are not content components; they are hiding content components like "number" and "street”

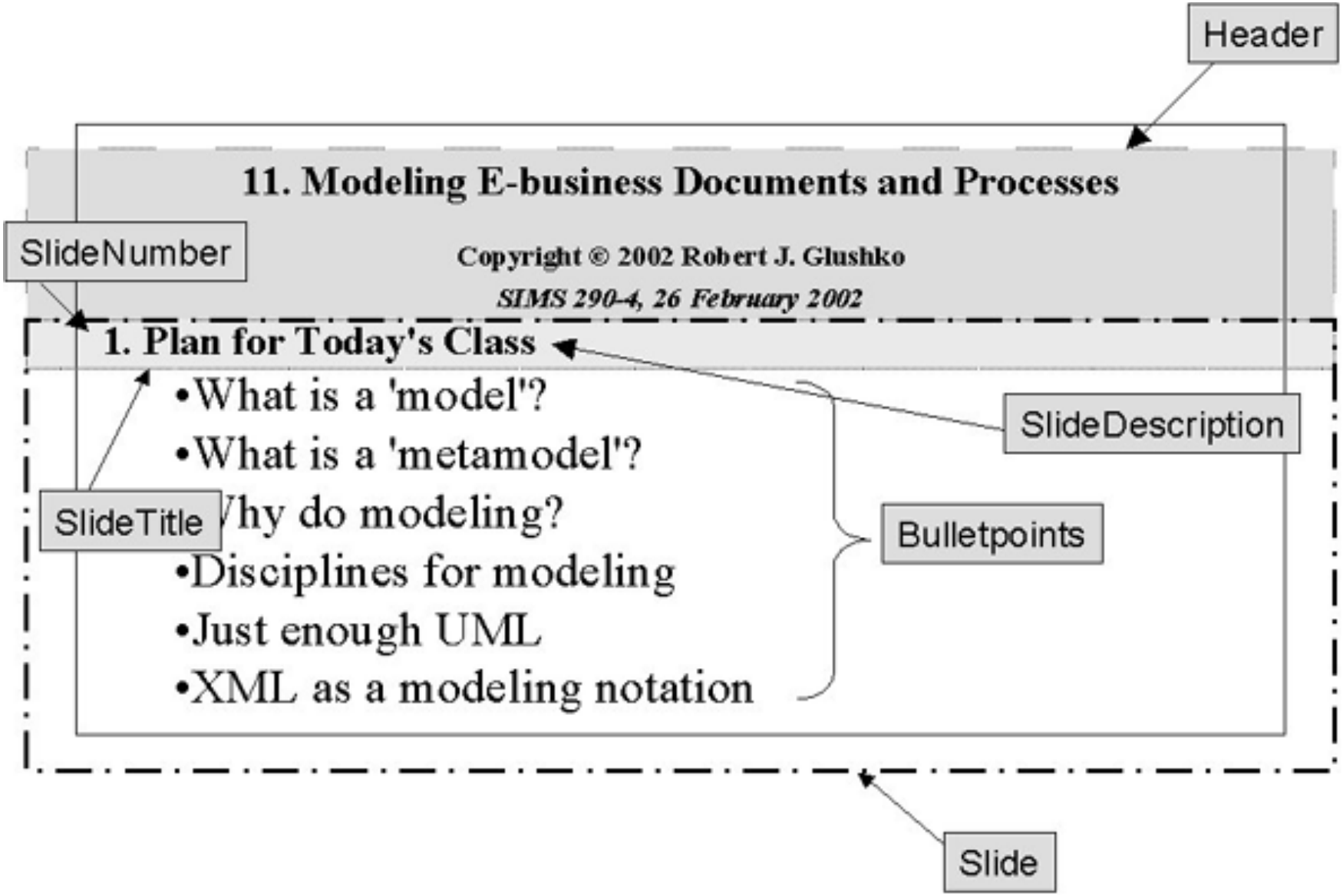
Presentation View of Lecture Slide



Structural Information

- Removing the presentation components allows us to focus on the two remaining types, Structure and Content.
- Structure is the physical or logical arrangement of components. Paragraphs, titles, sections, footnotes, tables, "parts" in a form, are all structural components.
- The structural components provide the hierarchical "skeleton" or "scaffold" into which the content components are arranged

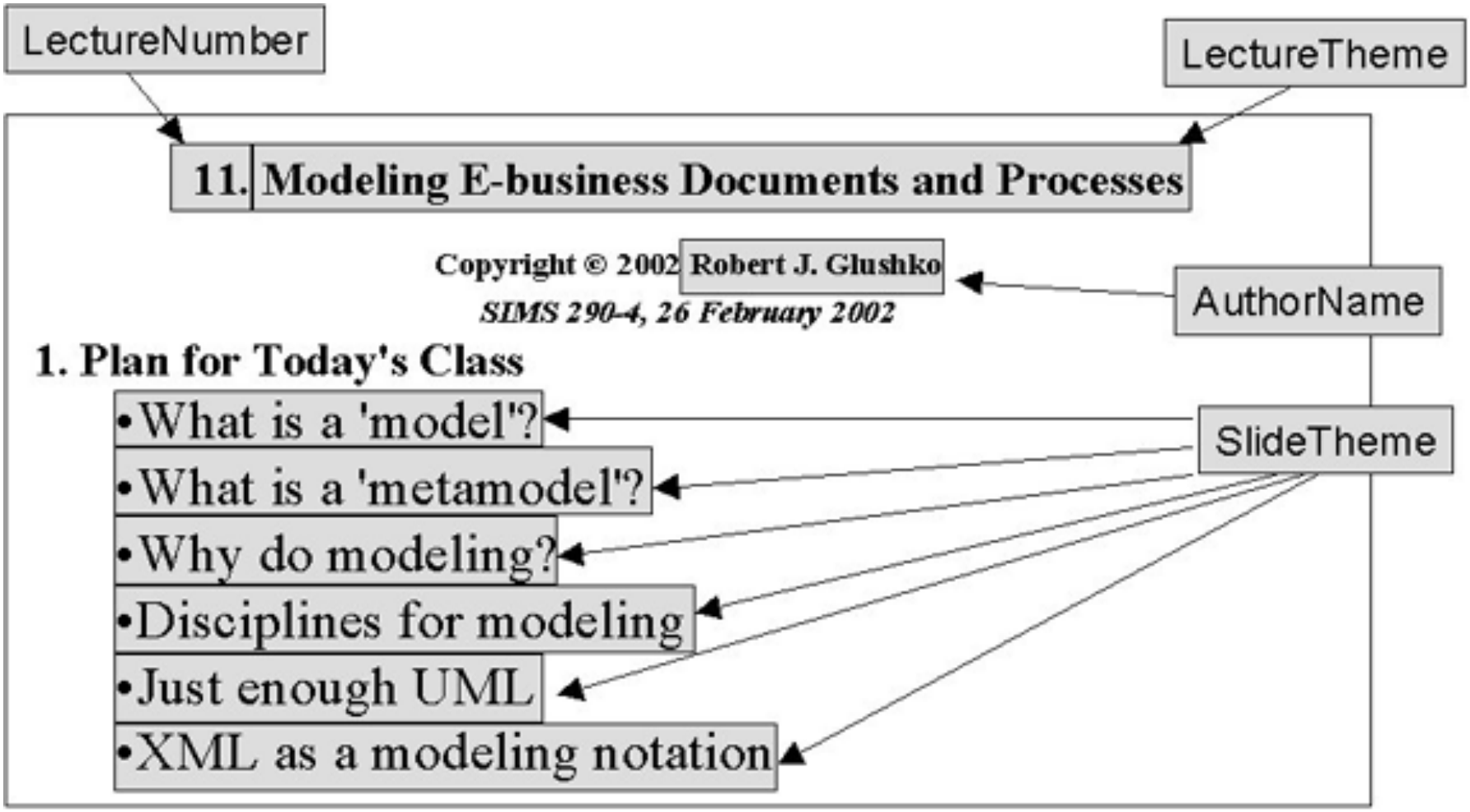
Structural View of Lecture Slide



Content Components

- The search for “meaning” in our document analysis ends up with the content.
- Content components are the "nouns" in our documents or sets of data – things like "topic," "summary," "name," "address," "price"

Content View of Lecture Slide



Deconstruct into Presentation, Structure, and Content

Rich and Simple French Onion Soup

We have been trying French onion soup in restaurants for years and my family and friends agree none can compare to my recipe for taste and simplicity of preparation. Makes 4 to 6 servings.

---

4 cups sliced onions
4 (10.5 ounce) cans beef broth
1/2 cup unsalted butter
2 tablespoons olive oil
2 tablespoons dry sherry (optional)
1 teaspoon dried thyme
salt and pepper to taste
1 (1 pound) loaf French bread, sliced
6 slices provolone cheese
3/4 cup diced Swiss cheese
1/4 cup grated Parmesan cheese

Directions

1 Melt butter in an 8 quart stock pot on medium heat. Add olive oil and stir. Add onions and continually stir until tender and translucent. Do not brown the onions.

2 Add beef stock, sherry and thyme. Season with salt and pepper, and simmer for 30 minutes.

3 Ladle soup into individual, oven safe, serving bowls and place one slice of bread on top, (it can also be broken into pieces, whichever you prefer). Layer cheese on top of bread; placing a slice of provolone,1/2 slice diced Swiss and then Parmesan cheese. Place bowls on cookie sheet and broil until cheese bubbles and browns slightly.

Generated or Derived Components

- "Table of Contents," "Permuted Index," and list of figures, tables, or other types of components can usually be generated or derived from other components and are not components in their own right
- Similarly, if “TotalPrice" is "Quantity" x "UnitPrice" we might only want the latter two components in our model since collecting that first one separately could lead to data integrity problems

Harvesting and Consolidation

- Harvesting – Create a set of candidate content components by extracting them from the information sources while removing presentation and structure
- Consolidation– Identify synonyms and homonyms among the candidate content components, assigning a unique name to each unique meaning as part of a controlled vocabulary

Harvesting Components from Multiple Sources

Information Elements

Analyzed Documents or Applications

	1	2	3	4	5	6	7	…
1	x	x	x	x	x	x	x	…
2	x	x	x			x		…
3	x	x		x	x		x	…
4	x		x	x				…
5	x	x	x	x	x	x	x	…
6		x		x				…
7	x		x				x	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

The Simplest Information Component Model

- The simplest or minimal information component model is a glossary – a list of the words used to describe or name the "things of significance" and what they mean
- This simple data model is augmented as attributes or characteristics of the significant things are identified and recorded
- The model is further developed as relationships or associations or links between the "significant things" are identified and recorded

Information About Candidate Components

- What attributes about each type of content might we record in our analysis?
  - Names/synonyms/homonyms (what it is called) and Identifiers
  - Definition (what it "means")
  - Optionality, Cardinality (occurrence rules)
  - Data Type (text, numbers, date, video) and possible values, code sets, defaults
  - Relationships/Associations
  - ???

Analyzing "Possible Values"

- It is critical to capture any rules governing the possible values for a component
- Sometimes possible values are conventional, fixed, and span the entire semantic range for some domain (days of week, AM/PM)
- Determine who can control the value sets (internal [Manufacturer part #s] vs external [Bar codes, AAT])
- Patterns like regular expressions are often useful but not sufficient for validation
- And if the set of possible values isn't well motivated, fix it in your component design

Consolidating the Harvest

- We can begin our consolidation with the candidate components from any of the information sources, but we recommend using the one you believe is the most authoritative or that yielded the most components
- The goal is to combine components that are synonyms (different names for the same meaning) and to distinguish any homonyms (same names for different meanings)
- It is desirable for a set of components to enable one and only one way to describe something

Merging Candidate Components

SOURCE 1 CANDIDATE COMPONENTS
A
B
C
D

SOURCE 2 CANDIDATE COMPONENTS
A
D
E (synonym of C)
F

SOURCE 3 CANDIDATE COMPONENTS
B (homonym of B in source 1)
C
D

Consolidating Candidate Components

CONSOLIDATED TABLE OF CONTENT COMPONENTS

Name	Semantic Description	Source 1	Source 2	Source 3
Title	The title of the event	X	X	
Start Date	The date of the event, or the first date of a recurring event	X		
End Date	The last date of the event	X		
Location	The location of the event	X	X (merged with synonym *Venue*)	X
Speaker	Name(s) of the person(s) speaking at the event		X	X
Description	The description of the event		X	X
Speaker Title	The title of the speaker			X (renamed homonym *Title*)

Figure 12-13. Completed Consolidated Table of Content Components

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

8 October 2013
Lecture 12.5 –
"Heavyweight" vs. "Lightweight" Modeling

Modeling Methodologies

- When we create a model we follow – implicitly or explicitly – a modeling methodology, the steps or techniques for analysis, design, and implementation
- Methodologies can be formal, prescriptive, step-by-step, documented with artifacts at each step, and auditable
- Or they can be the opposite: informal, ad hoc, "seat of the pants" with no trace other than the final model artifact itself

Heavyweight vs. Lightweight Modeling (1)

- At one end of the modeling continuum, “heavyweight” methods are based on consensus requirements that are identified and then satisfied in the most robust way possible
- These "heavyweight" approaches are expensive to follow but the models are credible and prescriptive and enable “traceability” of analysis and design decisions
- Examples of standards organizations that follow heavyweight methods: OASIS, W3C

Heavyweight vs. Lightweight Modeling (1)

- At the extreme "lightweight" end of the modeling continuum, document models tend to be inductive, descriptive, and "as is"
- They codify simple and commonly used information models (like "microformats")
- Schema.org contains a somewhat larger set of information models and they appear to have been created more systematically - but this is a proprietary effort with little transparency

There Are No Modeling Shortcuts

- You might think that "modeling" means "writing a schema given a set of instances" or "inferring a schema from a single instance" (like you can with the "autogenerate" function in many XML editors)
- But schemas developed without a stage of conceptual design (other than very simple ones) are rarely very useful because they are too closely tied to the particular instances used, which may not be representative
- Sometimes schemas went through a stage of conceptual design but once the schemas are implemented the conceptual information isn't available to allow users to evaluate suitability

Readings for Next Lecture

- TDO 8.2.1.6
- Graph Theory. Wikipedia. en.wikipedia.org/wiki/Graph_theory
- Kleinberg, Jon. "Analysis of large-scale social and information networks." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371, no. 1987 (2013). rsta.royalsocietypublishing.org/content/371/1987/20120378.short
- Krebs, V. Social Network Analysis: A Brief Introduction www.orgnet.com/sna.html
- MacRoberts, Michael H., and Barbara R. MacRoberts. "Problems of citation analysis." Scientometrics 36, no. 3 (1996): 435-444. (just the 12 problems… skim or skip "one paper: two philosophies")