10. Documents and Data Models... and Modeling

INFO 202 - 1 October 2008

Bob Glushko

Plan for INFO Lecture #10

Modeling across the "Document Type Spectrum"

Document models {and,or,vs} data models

"Berkeley Event Calendar Network" case study

How much modeling is necessary?

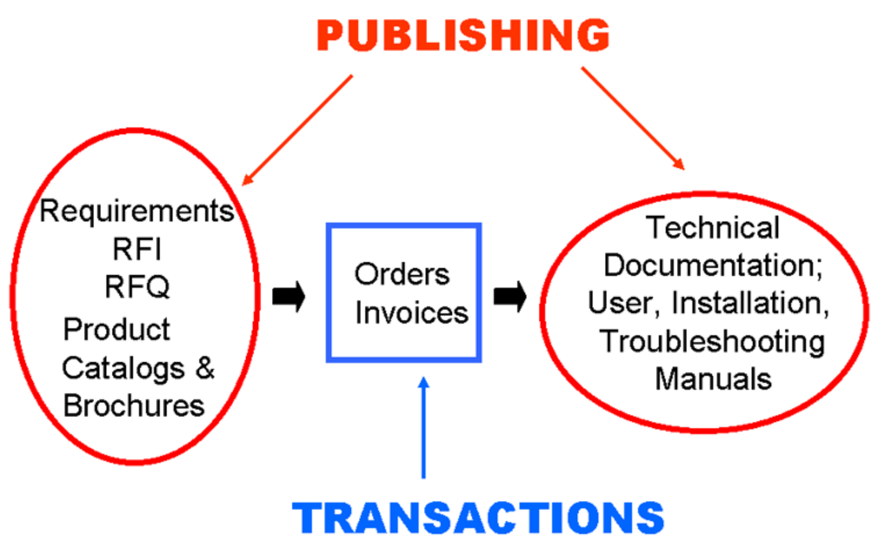
Documents vs. Data

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.

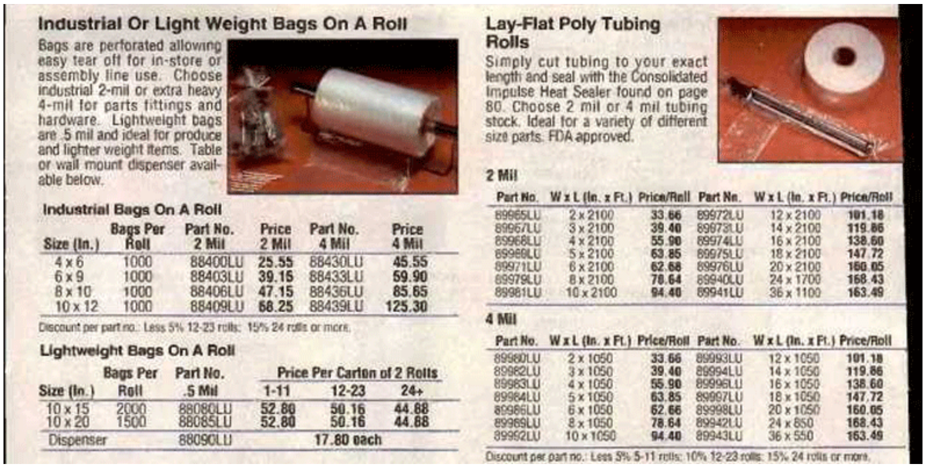
This document vs. data distinction is embedded and reinforced in courses, textbooks, technology, and product marketing

And it doesn't help

Mixing Data and Documents



Catalog: Data (Document)



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

	Bags Per	Part No.	Price Per Carton of 2 Rolls		
Size (In.)	Roll	.5 Mil	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

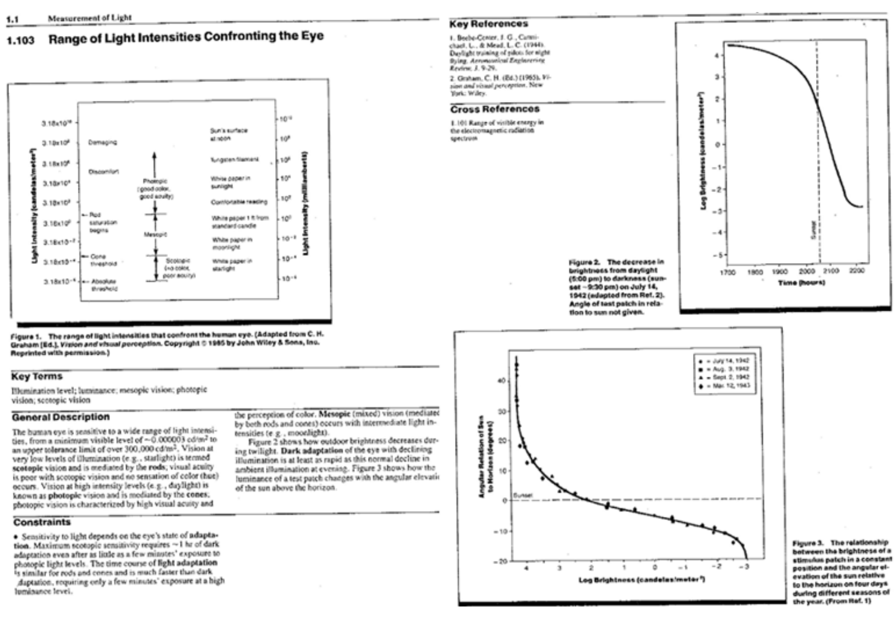
Part No.	W x L (In. x Ft.)	Price/Roll	Part No.	W x L (In. x Ft.)	Price/Roll
89965LU	2 x 2100	33.66	89972LU	12 x 2100	101.18
89967LU	3 x 2100	39.40	89973LU	14 x 2100	119.86
89968LU	4 x 2100	55.90	89974LU	16 x 2100	138.60
89969LU	5 x 2100	63.85	89975LU	18 x 2100	147.72
89971LU	6 x 2100	62.68	89976LU	20 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
89989LU	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.

Reference Book: Document (Data)



1.1 Measurement of Light

1.103 Range of Light Intensities Confronting the Eye

Figure 1. The range of light intensities that confront the human eye. (Adapted from C. H. Graham [Ed.], *Vision and visual perception*. Copyright © 1965 by John Wiley & Sons, Inc. Reprinted with permission.)

Key Terms

Illumination level; luminance; mesopic vision; photopic vision; scotopic vision

General Description

The human eye is sensitive to a wide range of light intensities, from a minimum visible level of ~0.000003 cd/m² to an upper tolerance limit of over 300,000 cd/m². Vision at very low levels of illumination (e.g., starlight) is termed **scotopic** vision and is mediated by the **rods**; visual acuity is poor with scotopic vision and no sensation of color (hue) occurs. Vision at high intensity levels (e.g., daylight) is known as **photopic** vision and is mediated by the **cones**; photopic vision is characterized by high visual acuity and the perception of color. **Mesopic** (mixed) vision (mediated by both rods and cones) occurs with intermediate light intensities (e.g., moonlight).

Figure 2 shows how outdoor brightness decreases during twilight. **Dark adaptation** of the eye with declining illumination is at least as rapid as this normal decline in ambient illumination at evening. Figure 3 shows how the luminance of a test patch changes with the angular elevation of the sun above the horizon.

Constraints

- Sensitivity to light depends on the eye's state of **adaptation**. Maximum scotopic sensitivity requires ~1 hr of dark adaptation even after as little as a few minutes' exposure to photopic light levels. The time course of **light adaptation** is similar for rods and cones and is much faster than dark adaptation, requiring only a few minutes' exposure at a high luminance level.

Key References

1. Beebe-Center, J. G., Carmichael, L., & Mead, L. C. (1944). Daylight training of pilots for night flying. *Aeronautical Engineering Review, 3,* 9-29.
2. Graham, C. H. (Ed.) (1965). *Vision and visual perception*. New York: Wiley.

Cross References

1.101 Range of visible energy in the electromagnetic radiation spectrum

Figure 2. The decrease in brightness from daylight (5:00 pm) to darkness (sunset ~9:30 pm) on July 14, 1942 (adapted from Ref. 2). Angle of test patch in relation to sun not given.

Figure 3. The relationship between the brightness of a stimulus patch in a constant position and the angular elevation of the sun relative to the horizon on four days during different seasons of the year. (From Ref. 1)

Contrasting Methodologies for Documents and Data

Documents and data have had two different disciplines or methods of analysis that have had little intersection

Document-centric analysis

Data-centric analysis

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

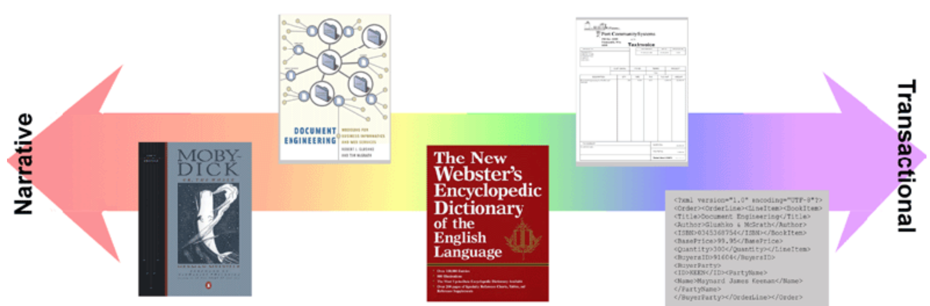
Data-Centric Analysis

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"

The Document Type Spectrum



The Document Type Spectrum – "Narrative Publications"

Authored by people

Highly designed, with rich presentational characteristics correlated with semantics and structure

Heterogeneous in structure and content

Weakly datatyped – "just text"

The Document Type Spectrum – "Transactional Documents"

Created mechanically

Few and somewhat arbitrary presentational characteristics

Homogeneous in structure and content

Strongly datatyped

It's Obviously 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

Crossing the Chasm with "Document Engineering" Methods

Document Engineering harmonizes the terminology 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 for information exchange

Harvesting and Consolidation

HARVESTING: Create a set of candidate content components by extracting them from the information sources while removing presentation and structure

- For each component, record its properties (or metadata or attributes or behaviors) that enable us to understand and distinguish it

CONSOLIDATION:(Identify synonyms and homonyms among the candidate content components, assigning a unique name to each unique meaning as part of a controlled vocabulary

- Names might follow precise rules to ensure that they can be reliably stored and located in a data dictionary (e.g., a la ISO 11179 part 5)

"Good Models and "Better Models"...

Definitions

Definitions in a controlled vocabulary

Data types

Metadata

Metamodels

Formal assertions

Ontologies and thesauri

The Simplest 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

Component Metadata

What attributes about each type of content might we record in our analysis?

- Names/synonyms/homonyms (what it is called)
- Definition (what it "means")
- Identifiers
- Cardinality/Optionality (occurrence rules)
- Restricted values, code sets, defaults
- Data Type (text, numbers, date, video)
- Relationships/Associations (participation in structures and "ontology")

Modeling "Events" for The Berkeley Calendar Network

The first published Document Engineering case study whose "snapshots" illustrate the analysis, modeling, and schema encoding approach

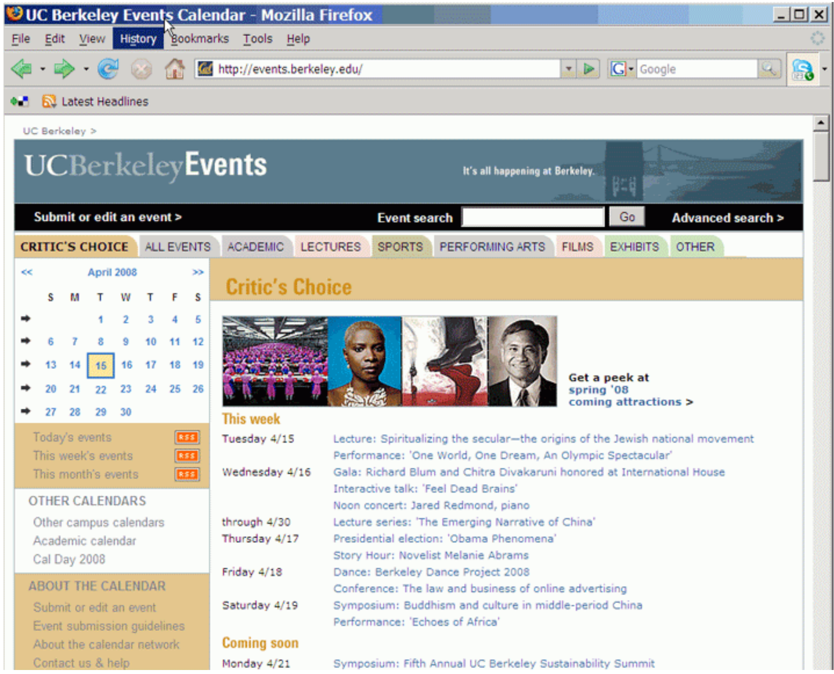
The problem - scores of calendars on berkeley.edu with overlapping coverage and audiences but incompatible data models

No automated reuse of information; you need to submit events to multiple calendars or copy events from them

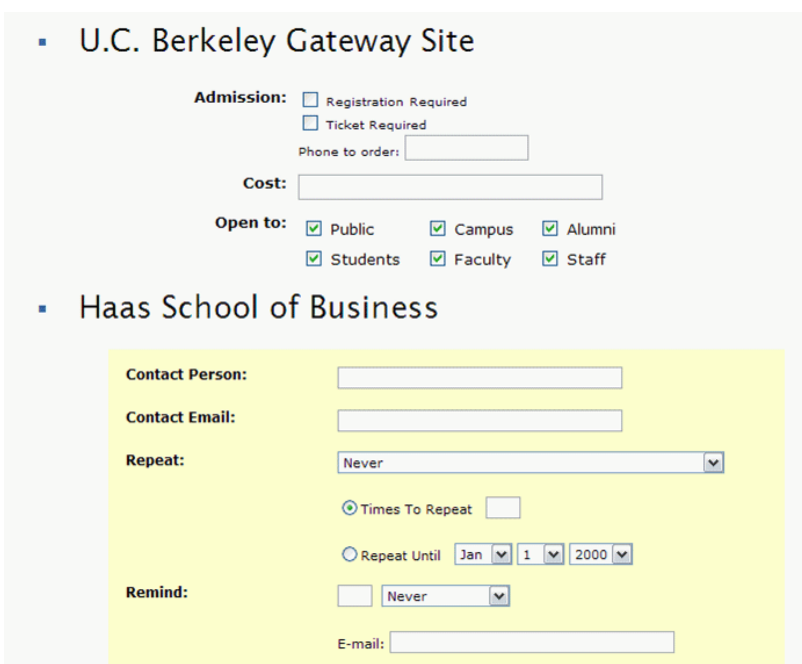
Each calendar has a different event submission form and a different model of an event

The UC Berkeley Event Calendar, 2004

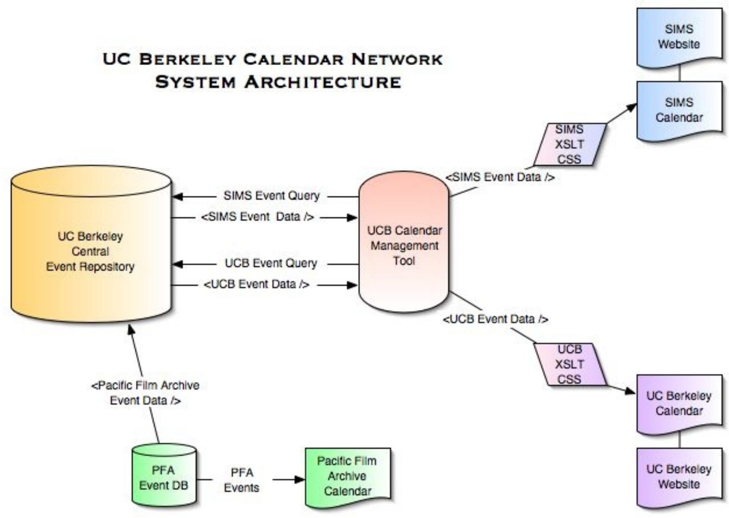
The UC Berkeley Event Calendar, 2008



Typical Incompatibility of Event Models



Event Calendar Network: Conceptual Architecture



Information Sources

User Interviews (18)

Event specifications/standards (iCalendar [IETF RFC 2445], SKICal)

Existing Calendars (23)

Event Calendars: Analysis Strategy

What can we learn from a specific calendar instance?

What can we learn from an "add new event" forms?

But you also have to look at instances and forms in combination

Kept analyzing new calendars until "law of diminishing returns" kicked in

Event Calendars: Harvesting and Consolidating Components

Synonyms:

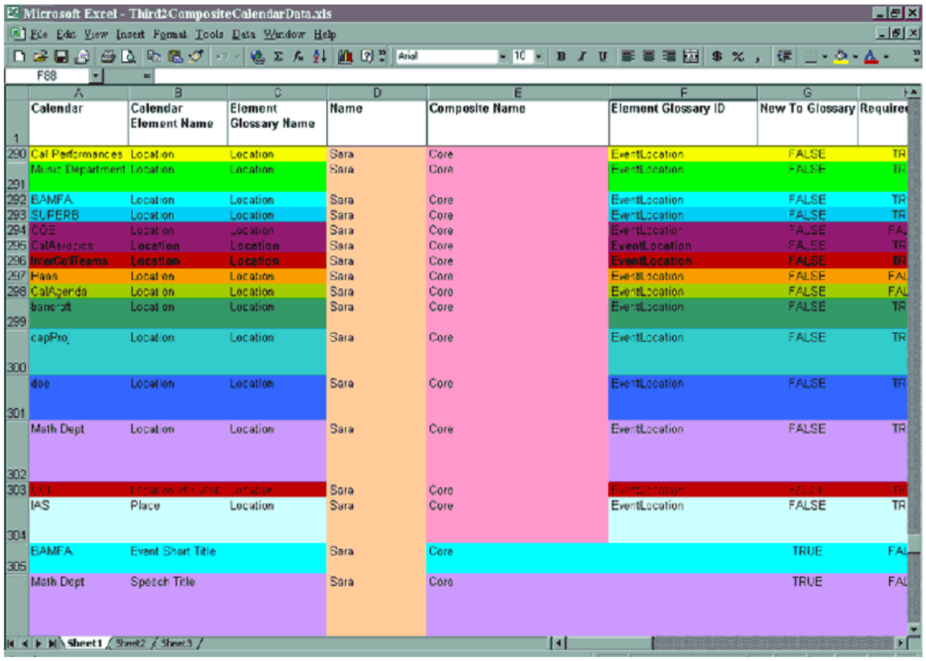
- Start Date
- Commencement

Homonyms:

- Contact (person submitting an event)
- Contact (person to contact about an event)
- Category / Type (disjoint domains: events, attendees)

Harvesting took on average 2 hours per calendar

Event Calendars Harvest of Candidate Components



Microsoft Excel - Third2CompositeCalendarData.xls

	Calendar	Calendar Element Name	Element Glossary Name	Name	Composite Name	Element Glossary ID	New To Glossary	Require
1								
290	Cal Performances	Location	Location	Sara	Core	EventLocation	FALSE	TR
291	Music Department	Location	Location	Sara	Core	EventLocation	FALSE	TR
292	BAMFA	Location	Location	Sara	Core	EventLocation	FALSE	TR
293	SUFERB	Location	Location	Sara	Core	EventLocation	FALSE	TR
294	COE	Location	Location	Sara	Core	EventLocation	FALSE	FA
295	CalAerobics	Location	Location	Sara	Core	EventLocation	FALSE	TR
296	InterCollTeams	Location	Location	Sara	Core	EventLocation	FALSE	TR
297	Haas	Location	Location	Sara	Core	EventLocation	FALSE	FAL
298	CalAgenda	Location	Location	Sara	Core	EventLocation	FALSE	FAL
299	bancroft	Location	Location	Sara	Core	EventLocation	FALSE	TR
300	capProj	Location	Location	Sara	Core	EventLocation	FALSE	TR
301	doe	Location	Location	Sara	Core	EventLocation	FALSE	TR
302	Math Dept	Location	Location	Sara	Core	EventLocation	FALSE	TR
303	[illegible]	[illegible]	[illegible]	Sara	Core	[illegible]	[illegible]	[illegible]
304	IAS	Place	Location	Sara	Core	EventLocation	FALSE	TR
305	BAMFA	Event Short Title		Sara	Core		TRUE	FAL
	Math Dept	Speech Title		Sara	Core		TRUE	FAL

Sheet1 / Sheet2 / Sheet3

Event Calendars Component Consolidation (Simplified)

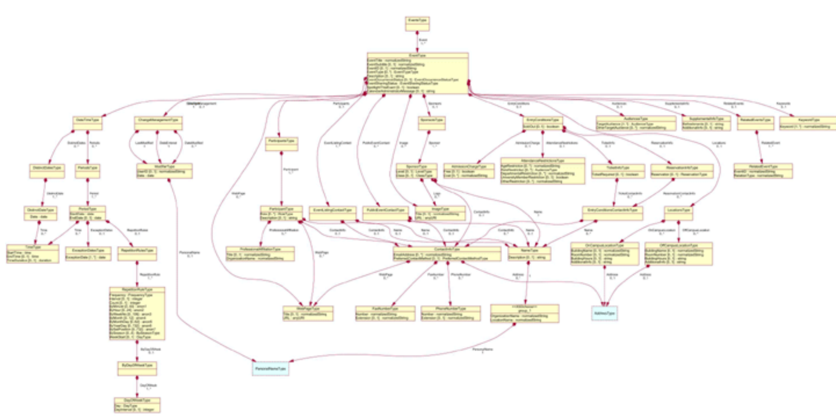
Name	Semantic Description	Source 1	Source 2	Source 3
Title	The title of the event	X	X	
Start Date	The date of the event, or the first date of a recurring event	X		
End Date	The last date of the event	X		
Location	The location of the event	X	X (merged with synonym *Venue*)	X
Speaker	Name(s) of the person(s)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

Event Calendars: The Conceptual Model

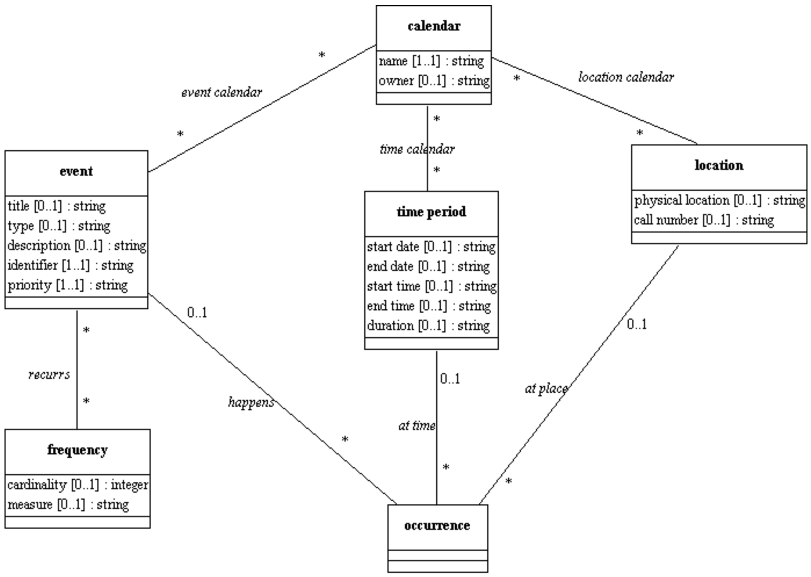
When we've analyzed all of the candidate components for dependencies, we've created a conceptual model for event calendars

From this model we can assemble any of a set of related document types for different varieties of event calendars

The Complete Conceptual Model



A Simplified Conceptual Model



Document Models {and,or,vs} Data Models

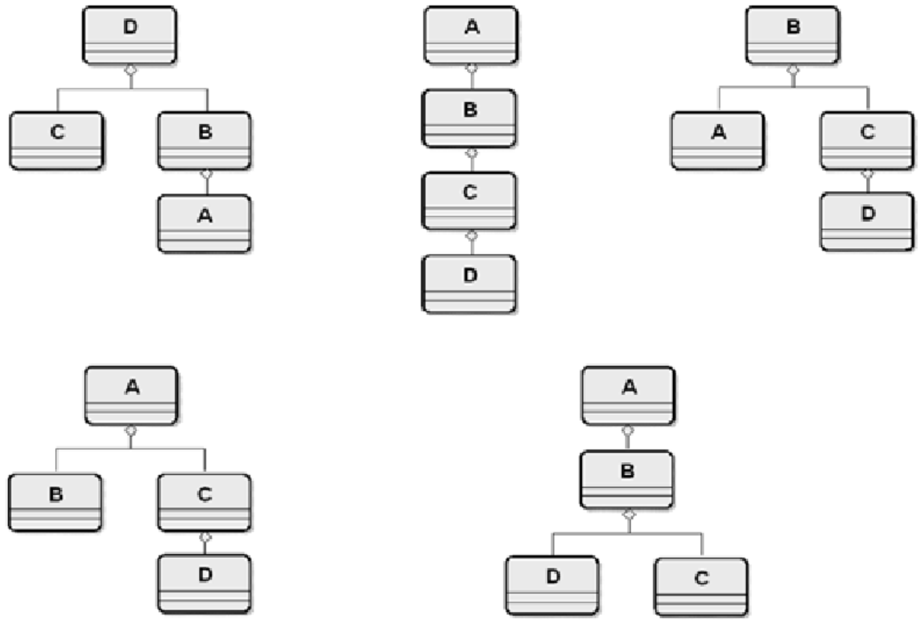
A relational model (a set of tables in our example) simultaneously describes all of the associations among the components; put another way, it doesn't highlight any particular association

But when we exchange information, we do so to satisfy the requirements in some context

If there are multiple ways to interpret the content we will not achieve interoperability

So we impose a contextual interpretation when we create a hierarchy on a relational model

Multiple Paths in the Component Network



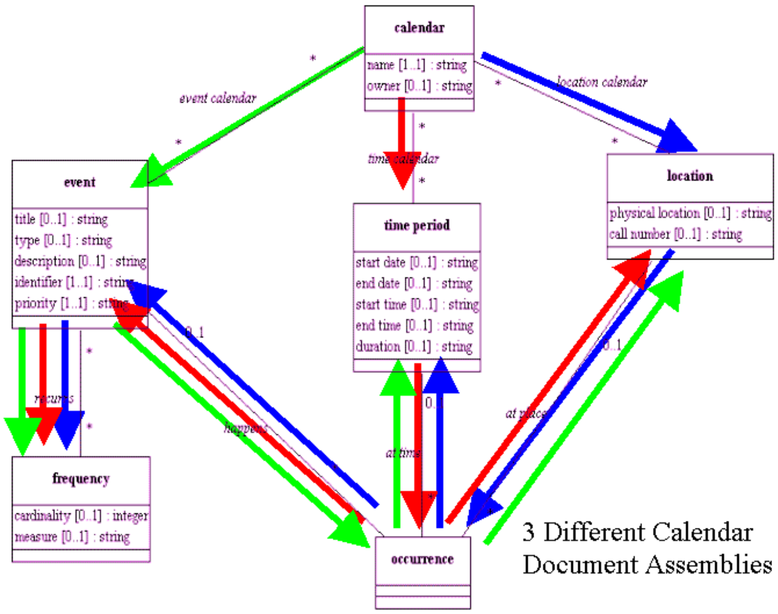
Document Model Assembly

Document model assembly is the process of creating a model of a document type – hierarchical and nested – by drawing on the "pool" or library of content and structural components

Assembly involves designing (or selecting a pattern for) the top level structure as an entry point and then navigating through the relationships in the conceptual model to order components to satisfy requirements

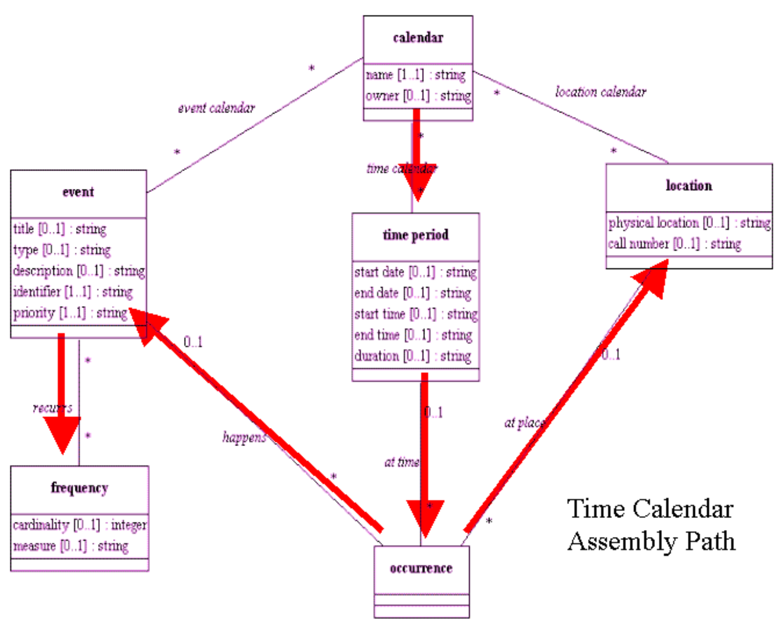
Assembly order can differ whenever there is a bi-directional relationship between components

Alternate Assemblies



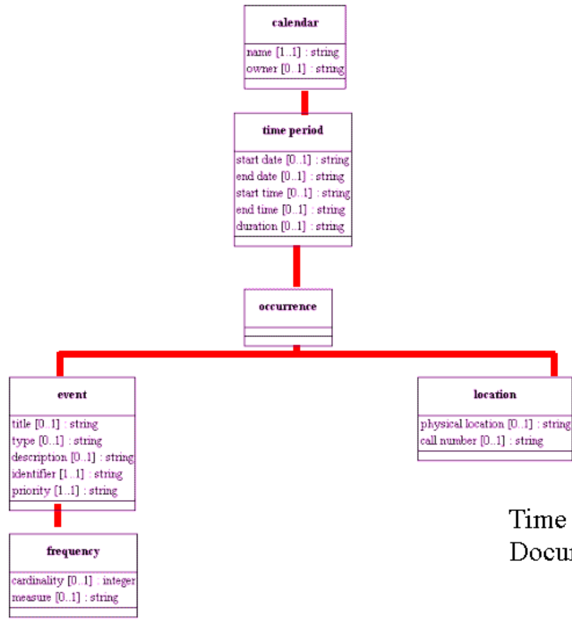
3 Different Calendar Document Assemblies

Assembling a Time-based Calendar Model



Time Calendar Assembly Path

The Time-Based Calendar Model



Time Calendar
Document Model

The Modeling Debate [1]

Some problems and some domains are inherently complex and a careful, rigorous modeling approach is required

- This "heavyweight" position argues that there are no modeling shortcuts

But some people argue that modeling "involves a substantial amount of work that is often political, tedious, and unpleasant" that should be avoided whenever possible

- Some domains and use cases might be simple enough ("Microformats") that less "heavyweight" modeling approaches could suffice

The Modeling Debate [2]

You should always look to see if someone has already modeled your problem domain (Cover Pages and OASIS)

If the underlying conceptual model of an existing vocabulary doesn't fit your requirements and you must develop your own, you have many choices to make about scope, abstraction, and granularity

Modeling "Professor Stories"

Bob Glushko is an Adjunct Full Professor at UC Berkeley's School of Information, located in South Hall. He teaches Information Organization and Retrieval (INFO 202), Document Engineering (INFO 243), and other courses. He has a B.A. from Stanford University (California) and a Ph.D. from UC San Diego.

Coye Cheshire is an Assistant Professor at the School of Information. He recently received his Ph.D. from Stanford University. He teaches Computer Mediation Communication, Social and Organizational Aspects of Computing, and other courses.

Modeling as "Text Blobs"

<Para> Bob Glushko is an Adjunct Full Professor at UC Berkeley's School of Information, located in South Hall. He teaches Information Organization and Retrieval (INFO 202), Document Engineering (INFO 243), and other courses. He has a B.A. from Stanford University (California) and a Ph.D. from UC San Diego. </Para>

"Content Nuggets" in the Text (aka "Mixed Content")

<Para><Name>Bob Glushko</Name> is an <Rank>Adjunct Full Professor</Rank> at the <Institution>UC Berkeley</Institution> <AcademicUnit>School of Information</AcademicUnit>, located in <Building>South Hall</Building>.
He teaches <CourseName>Information Organization and Retrieval</CourseName> (<CourseNum>INFO 202</CourseNum>), <CourseName>Document Engineering </CourseName> (<CourseNum>INFO 243</CourseNum>), and other courses.
He has a <Degree>B.A</Degree> from <Institution>Stanford University</Institution> (<State>California</State>) and a <Degree>Ph.D.</Degree> from <Institution>UC San Diego</Institution></Para>

A More Structured Professor Story

```
<Professor>
<Name>Bob Glushko</Name>
<Rank>Adjunct Full Professor</Rank>
<Affiliation>
        <Institution>UC Berkeley</Institution>
        <AcademicUnit>
                <Name>School of Information</Name>
                <Building>South Hall</Building>
        </AcademicUnit>
</Affiliation>
<Courses>
        <Course>
                <Name>Information Organization and Retrieval</Name>
                <Number>INFO 202</Number>
        </Course>
        ...
</Courses>
<Degrees>
        <Degree>
                <Institution State="California">Stanford University</Institution>
                <Type>B.A.</Type>
        </Degree>
        ...
</Degrees>
</Professor>
```

Facts in Tabular Format

NAME	RANK	AFFILIATION	COURSES	DEGREES
Bob Glushko	Adjunct Full Professor	UC Berkeley, School of Information (South Hall)	Info Org & Retrieval (INFO 202); Document Engineering (INFO 243)	BA, Stanford (California); Ph.D. UCSD
Coye Cheshire	Assistant Professor	UC Berkeley, School of Information	Computer-Mediated Communication; Social & Organizational Aspects of Computing	Ph.D, Stanford

Problems with this Organization of the Facts

It may seem that this way of organizing the facts is useful, but there are some problems with it

This is a "spreadsheet" style of data organization, with rows and columns defining cells that are just "data buckets" buckets into which we can put almost anything

Some of the "buckets" contain repeating items rather than "atomic" information components

Some of the "buckets" contain values that are not of the same type

What relationships describe how different columns go together?

Normalized Tables

PROFESSORS		
NAME	RANK	AFFILIATION
Bob Glushko	Adjunct Full Professor	UC Berkeley, School of Information (South Hall)
Coye Cheshire	Assistant Professor	UC Berkeley, School of Information

COURSES		
INSTRUCTOR	COURSE NAME	COURSE NUMBER
Bob Glushko	Info Org & Retrieval	INFO 202
Bob Glushko	Document Engineering	INFO 242
Coye Cheshire	Computer-Mediated Communication	
Coye Cheshire	Social & Organizational Aspects of Computing	

DEGREES		
RECIPIENT	DEGREE TYPE	INSTITUTION
Bob Glushko	B.A.	Stanford
Bob Glushko	Ph.D.	UCSD
Coye Cheshire	Ph.D.	Stanford

Readings for INFO Lecture #11

Robert J. Glushko and Tim McGrath, Document Engineering, Chapter 6, "When Models Don't Match: The Interoperability Challenge"

Michael Stonebraker and Joseph Hellerstein, "Content Integration for E-Business"