Alternatives for Representing Coding of Qualitative Data in DDI

Larry Hoyle

Institute for Policy & Social Research

University of Kansas

Qualitative Data

- Digital or real-world object (analog?)
- - Purpose collected, gathered, or referenced

  - Text
  - XML
  - Web pages
  - Video
  - Audio
  - Images

Three Scenarios

1. Documentation at the object level
2. Segments of objects need documentation

   E.g. CAQDAS – codes associated with defined segments
3. Segments have documentation and quantitative data have been generated from the qualitative segments

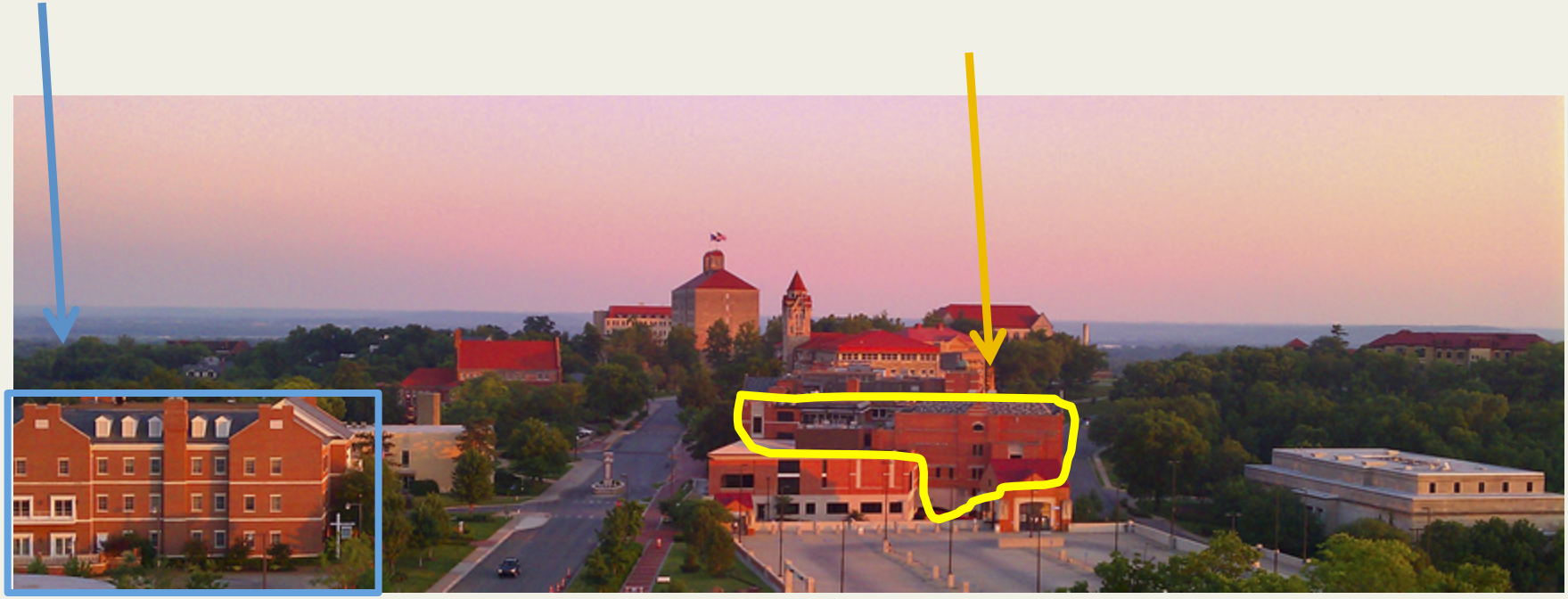
   E.g. text mining

DDI Qualitative Data Model Working Group http://www.ddialliance.org/alliance/working-groups#qdewg

Segments Examples

Segment defined by a rectangle

Segment defined by a polygon

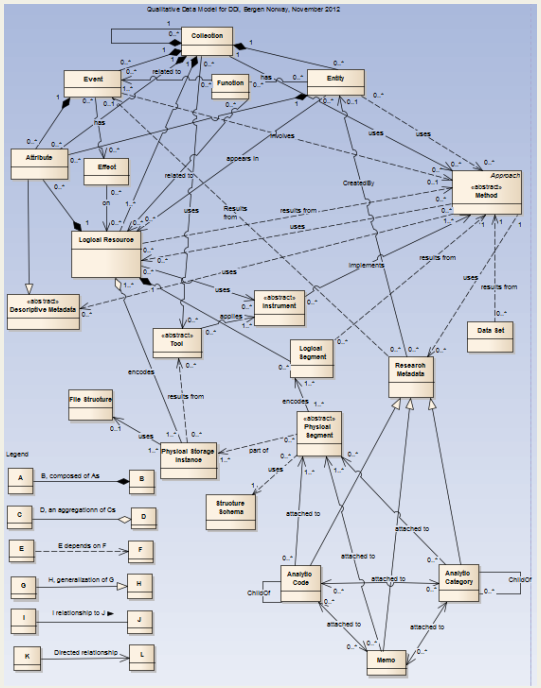


Segments can also be marked in text. Like in this text example.

Text segment

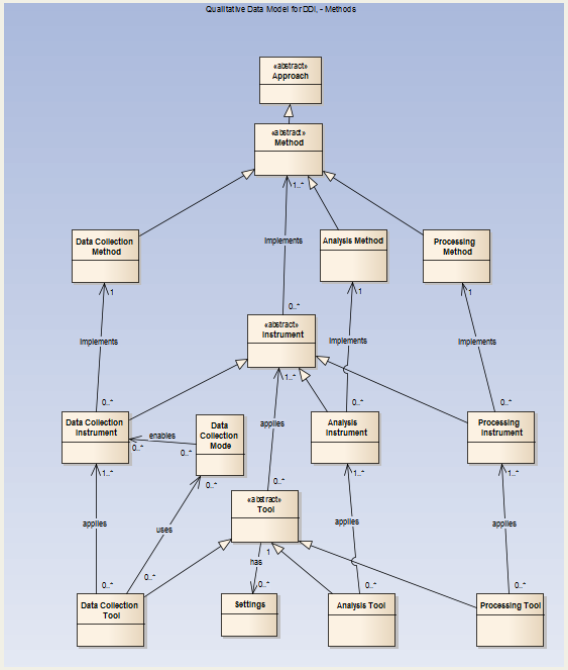
Overlapping text segment

Current Model

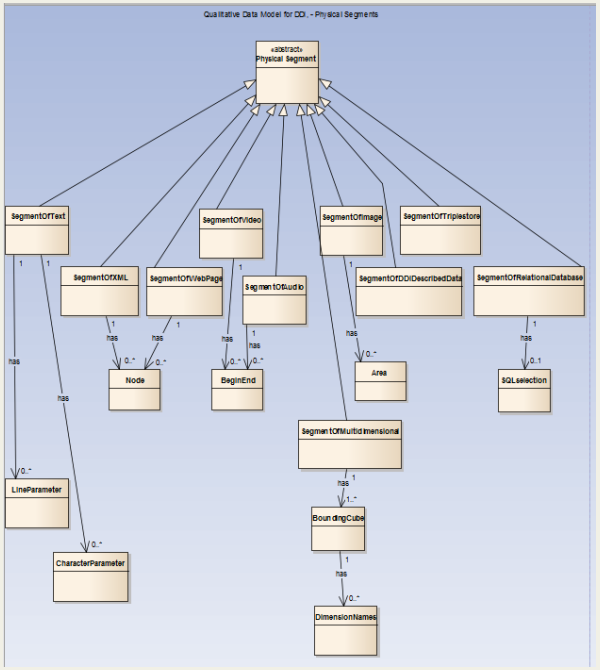


Overall model

(many current DDI elements assumed to be applicable and not shown)

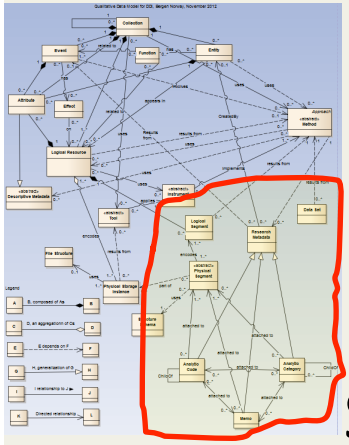


Methods and Instruments



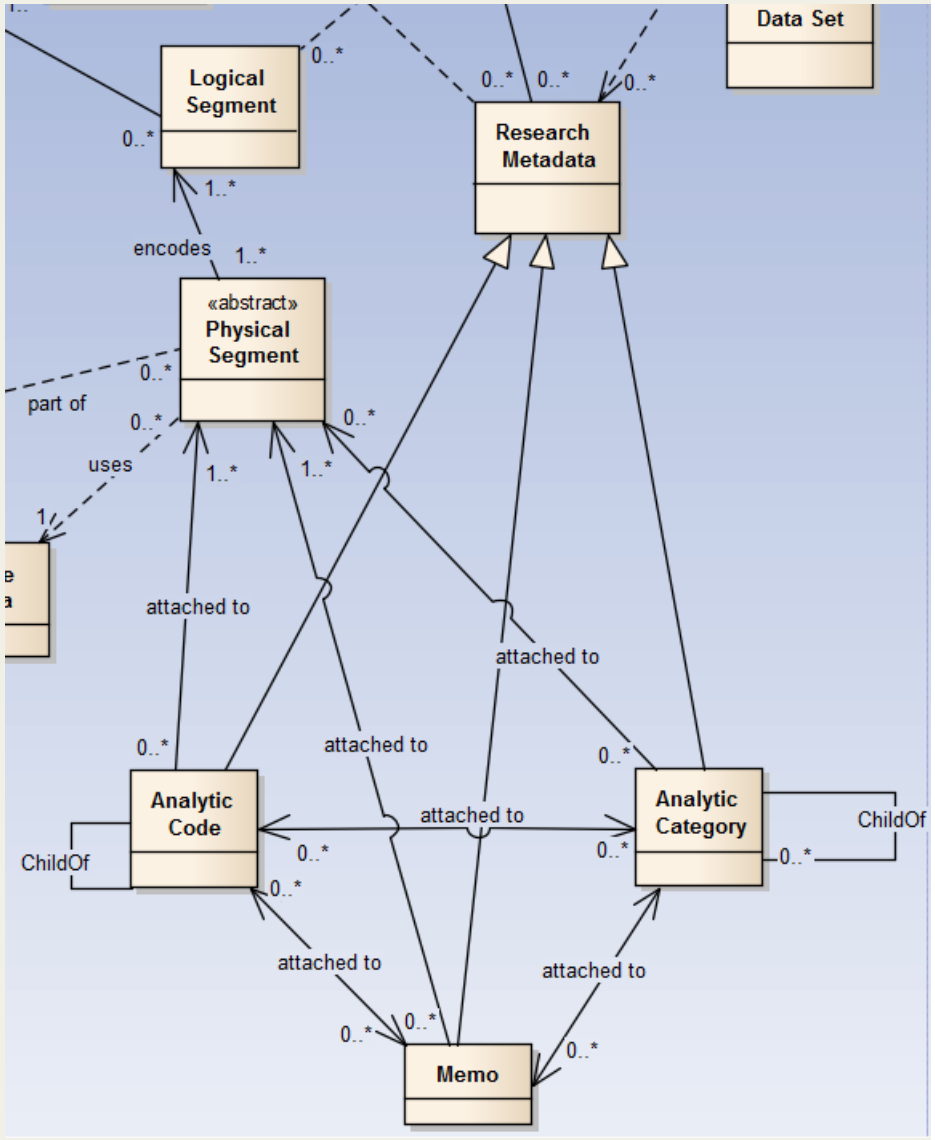
Segment Definition

Codes, Categories and Memos



Segments can have “Codes”, “Categories” and “Memos”

Modeled after terms from qualitative data analysis packages like Atlas/ ti or NVIVO

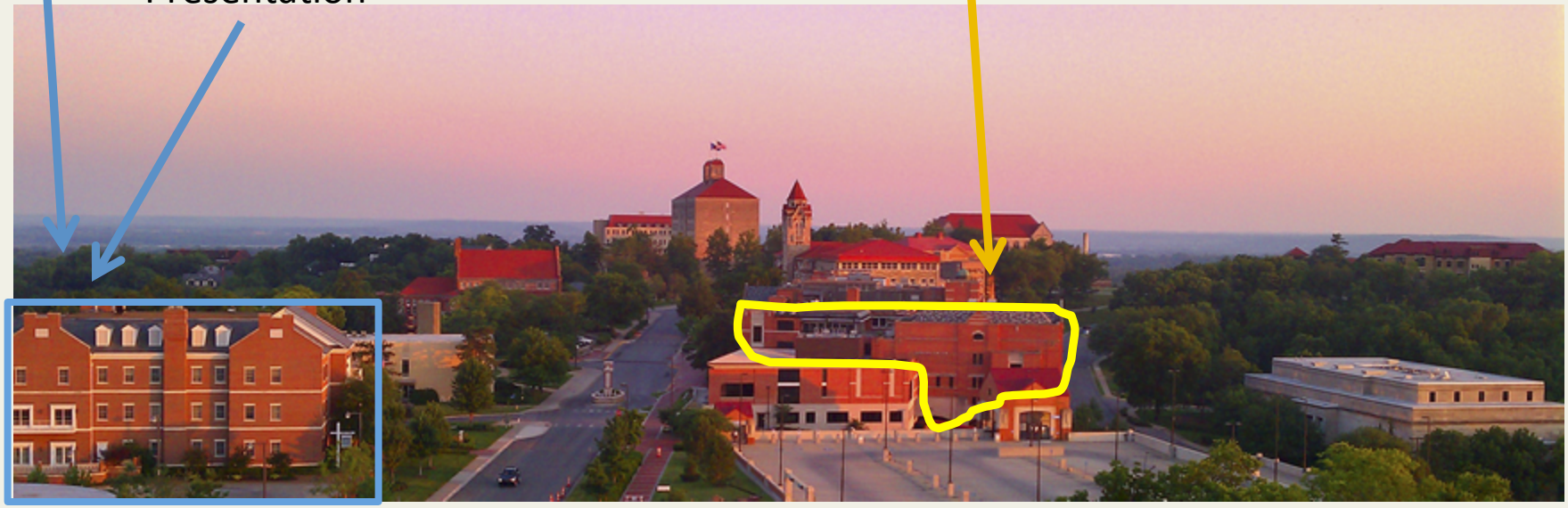


Codes and Memos

"Dinner site"

"Presentations"

Memo: This is marked here for an example in a PowerPoint Presentation

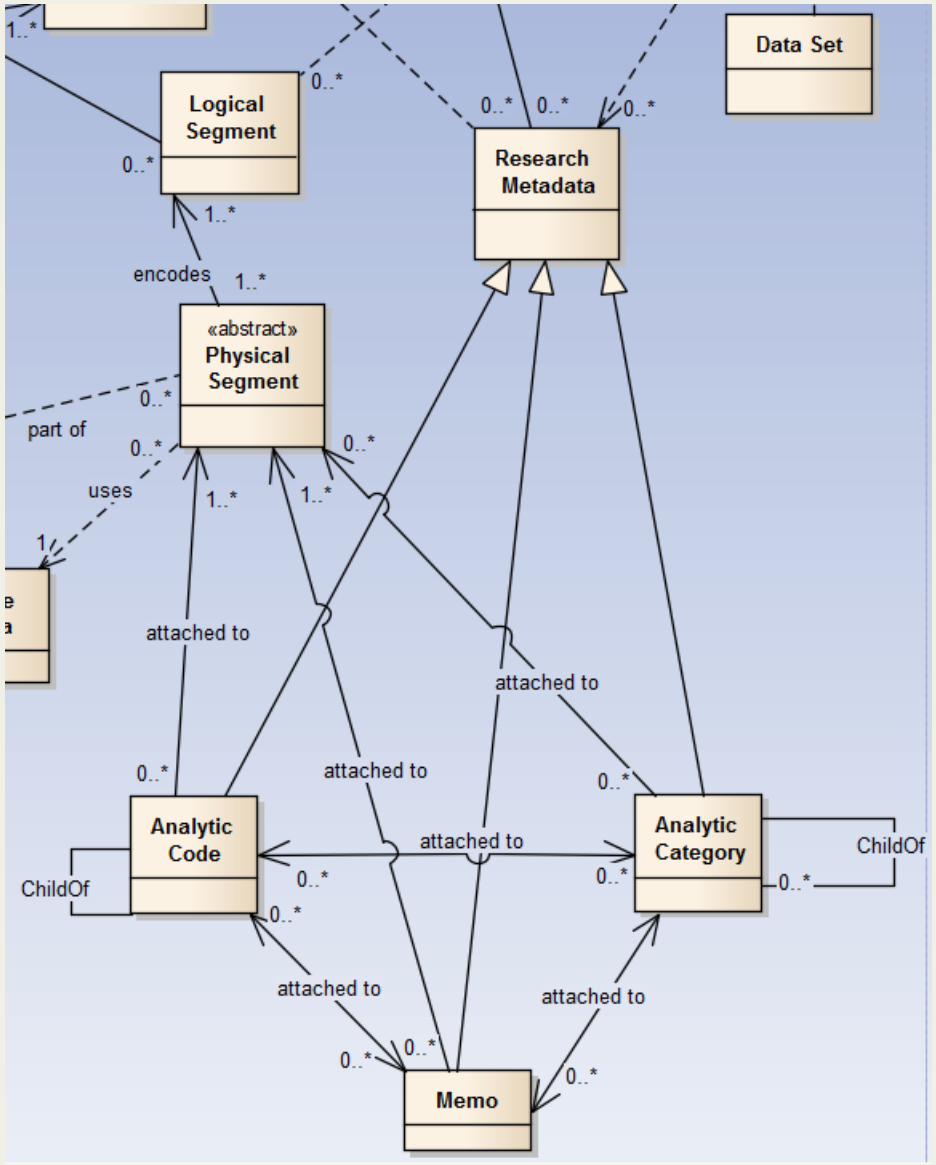


Segments can also be marked in text. Like in this text example.

"Marking"

"Text reference"

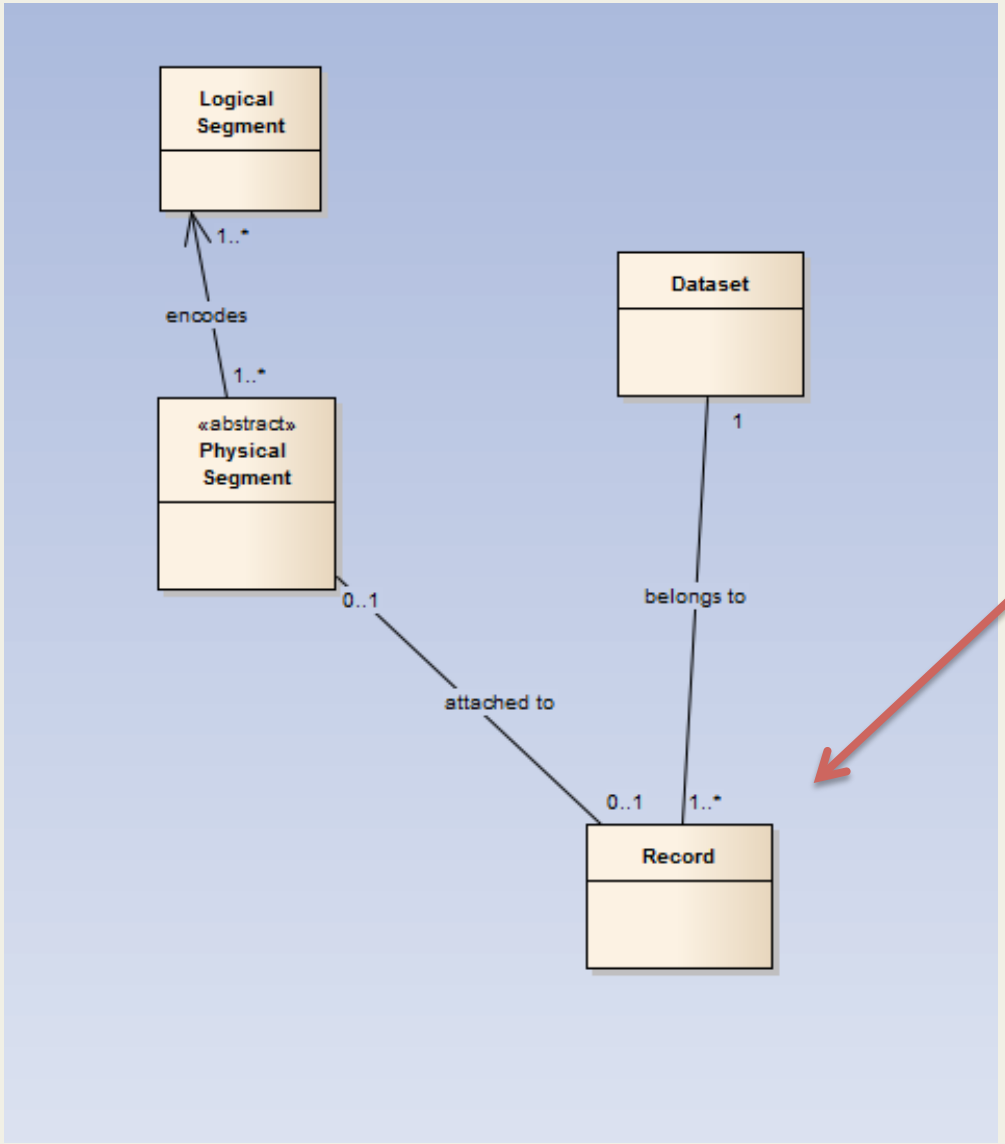
Dataset



Off by itself is this Dataset

This might be produced by text mining, or might be quantitative data associated with an open-ended question

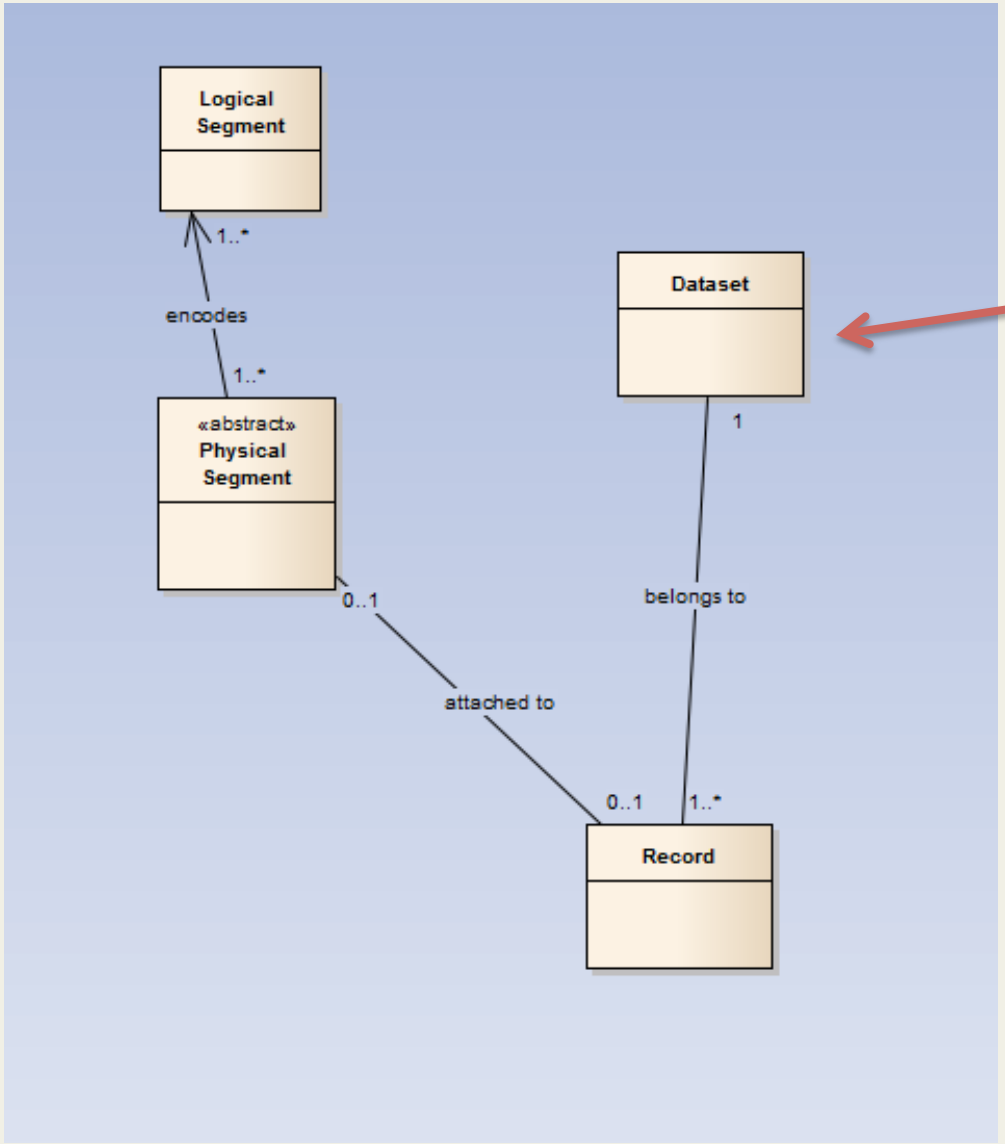
Alternative – Data/Metadata Record



This is a data record which can be described by existing DDI elements

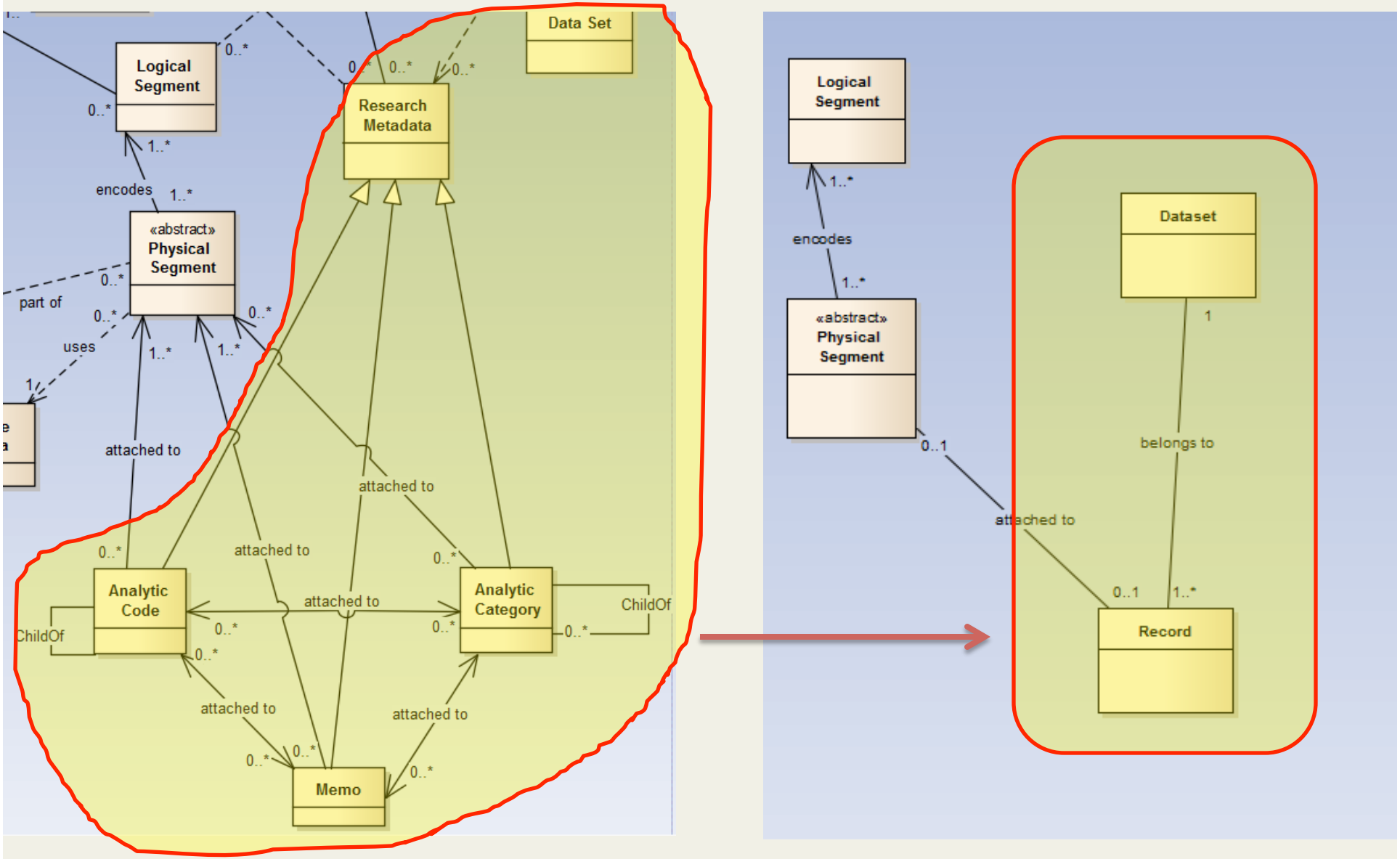
It can contain codes, categories, memos, and any quantitative data associated with the segment

Alternative - Dataset

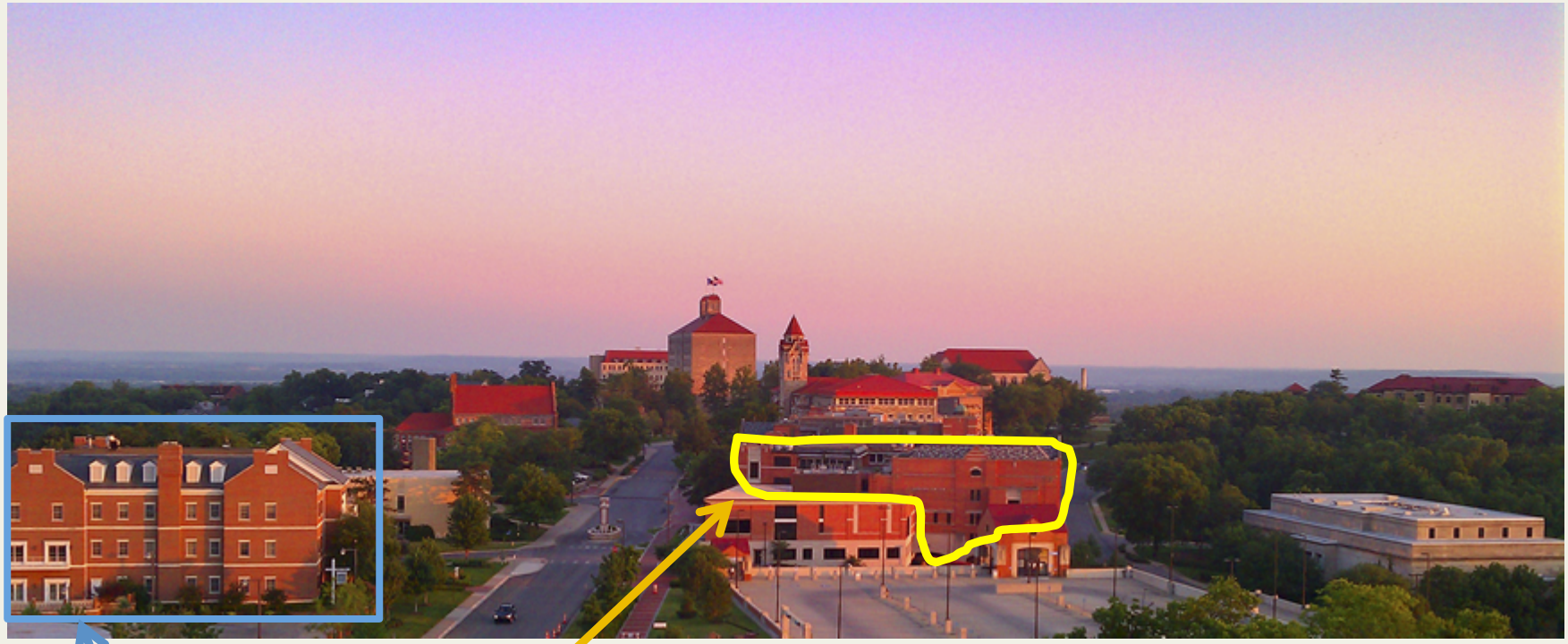


This dataset can be included in the DDI representation

Comparison


Logical Segment
Data Set
Research Metadata
encodes
«abstract» Physical Segment
part of
uses
attached to
Analytic Code
Analytic Category
ChildOf
Memo
Logical Segment
encodes
«abstract» Physical Segment
Dataset
belongs to
attached to
Record

Example Records



Dataset

SegmentID	SegmentName	CampusBuilding	EveningVenue	YearBuilt	Address	Cluster1	MiningVar
1	AlumniCenter	0	1	1983	1266 Oread Ave.	0	1.27
2	StudentUnion	1	0	1926	1301 JayHawk Blvd.	1	0.8

Example Record - Codes

Codes are handled like "Select all that apply" questions.

Codes

1=has code
0=does not

SegmentID	SegmentName	CampusBuilding	EveningVenue	YearBuilt	Address	Cluster1	MiningVar
1	AlumniCenter	0	1	1983	1266 Oread Ave.	0	1.27
2	StudentUnion	1	0	1926	1301 Jayhawk Blvd.	1	0.8

Could have Categories and Memo here as well

Example Record – Other Variables

Codes

Quantitative Variables

Text Mining Variables

SegmentID	SegmentName	CampusBuilding	EveningVenue	YearBuilt	Address	Cluster1	MiningVar
1	AlumniCenter	0	1	1983	1266 Oread Ave.	0	1.27
2	StudentUnion	1	0	1926	1301 Jayhawk Blvd.	1	0.8

Searching

Search for Evening Venue=1

SegmentID	SegmentName	CampusBuilding	EveningVenue	YearBuilt	Address	Cluster1	MiningVar
1	AlumniCenter	0	1	1983	1266 Oread Ave.	0	1.27
2	StudentUnion	1	0	1926	1301 Jayhawk Blvd.	1	0.8

Searching

Search for YearBuilt<1950

SegmentID	SegmentName	CampusBuilding	EveningVenue	YearBuilt	Address	Cluster1	MiningVar
1	AlumniCenter	0	1	1983	1266 Oread Ave.	0	1.27
2	StudentUnion	1	0	1926	1301 Jayhawk Blvd.	1	0.8

Text Mining Example

	title	paranum	storynum	paragraph
763	THE TALE OF JEMIMA PUDDLE-...	763	11	WHAT a funny sight it is to see a brood of ducklings with a hen! --List...
764	THE TALE OF JEMIMA PUDDLE-...	764	11	HER sister-in-law, Mrs. Rebeccah Puddle-duck, was perfectly willing...
765	THE TALE OF JEMIMA PUDDLE-...	765	11	"I wish to hatch my own eggs; I will hatch them all by myself," quacke...
766	THE TALE OF JEMIMA PUDDLE-...	766	11	SHE tried to hide her eggs; but they were always found and carried o...
767	THE TALE OF JEMIMA PUDDLE-...	767	11	Jemima Puddle-duck became quite desperate. She determined to m...
768	THE TALE OF JEMIMA PUDDLE-...	768	11	SHE set off on a fine spring afternoon along the cart- road that leads...
769	THE TALE OF JEMIMA PUDDLE-...	769	11	She was wearing a shawl and a poke bonnet.

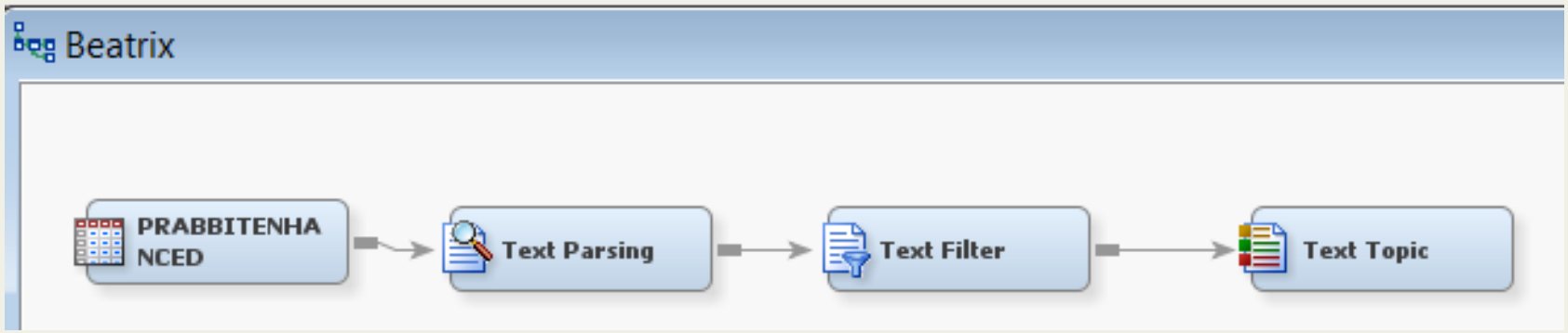
What a funny sight it is to see a brood of ducklings with a hen!

—Listen to the story of Jemima Puddle-duck, who was annoyed because the farmer's wife would not let her hatch her own eggs.

Each segment is a paragraph from a Beatrix Potter Story

(downloaded from Project Gutenberg - http://www.gutenberg.org/)

A Text Topic Tool Can Build a Set of Topics



Topics are Based on Weighted Combinations of Words

Topic	Category	Term Cutoff	Document Cutoff	
pigling,bland,piperson,alexander,+hen	Multiple	**0.173**	**0.419**	3

The weights for calculation

Terms

Topic Weight	+	Term	Role	# Docs	Freq
2.171		pigling	Prop	67	90
1.573		bland	Prop	38	46
0.644		alexander	Prop	17	23
0.305	+	paper	Noun	16	18
0.199	+	drop	Verb	11	12
0.173	+	wish	Verb	10	10

Each segment can then be assigned a score for that topic

Documents

Topic Weight	paragraph
2.234	"But I wish to preserve them for emergencies," said Pigling Bland
2.171	Pigling Bland drew forward a coppy stool, and sat on the edge of it,
2.106	Mr. Piperson poured out three platefuls: for himself, for Pigling, and
2.025	Pigling Bland slept like a top. In the morning Mr. Piperson made
1.876	"Now Pigling Bland, son Pigling Bland, you must go to market. Take

Future Segments Could be Scored on These

The weights for calculation

Terms

Topic Weight	+	Term	Role	# Docs	Freq
2.171		pigling	Prop	67	90
1.573		bland	Prop	38	46
0.644		alexander	Prop	17	23
0.305	+	paper	Noun	16	18
0.199	+	drop	Verb	11	12
0.173	+	wish	Verb	10	10

Described by a DDI Generation Instruction for a variable shared across studies?

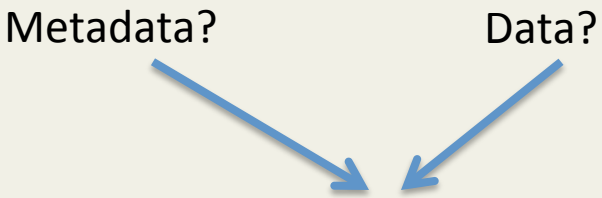
SegmentID	paragraph	MiningVar
1001	Did you read the paper?	0.305
1002	I wish spring would come	0.173

Data or Metadata?

Depends on use, doesn't it? (c.f. NSA use of phone "metadata")

YearBuilt might be metadata when searching for image segments or text descriptions of old buildings

It might be data if we used the text mining variables to predict building age



SegmentID	SegmentName	CampusBuilding	EveningVenue	YearBuilt	Address	Cluster1	MiningVar
1	AlumniCenter	0	1	1983	1266 Oread Ave.	0	1.27
2	StudentUnion	1	0	1926	1301 Jayhawk Blvd.	1	0.8

Similarity Between Models

- A code (its associated category) can refer to a concept
- A variable can refer to a concept

- So both approaches ultimately relate a segment to some concept

Data Record Advantages

- Can handle Mixed Method Research (Quantitative and Qualitative Approaches)
- Works for surveys with open-ended questions
- Allows for sharing of codes and quantitative variables across studies
- Searching for codes is the same as searching for other attributes (e.g. building age from above)
- Flexible

Disadvantages

- More difficult to explain?
- Searching involves more indirect lookup
  - Segment with a record with a value of 1 ("has attribute") on Variable "AnalyticCode"

  Vs

  - Segment with "AnalyticCode" of xxx
- "Codes", "categories", and "memos" lose special meaning
- Resistance from qualitative only researchers?

Discussion

- Advantages and disadvantages to data record approach?
- Other comments?

For More About the Qualitative Model

- See:
- ***Toward Qualitative Data in DDI*.** Larry Hoyle and Joachim Wackerow
- Larry Hoyle is a Senior Scientist at the Institute for Policy & Social Research, University of Kansas
- LarryHoyle@ku.edu .
- Joachim Wackerow is a metadata expert at GESIS - Leibniz Institute for the Social Sciences and can be reached at joachim.wackerow@gesis.org.

version: November 2013