Semi-structured data

• Data is not just text, but is not as well-structured as data in databases
• Occurs often in web databanks
• Occurs often in integration of databanks

Semi-structured data - properties

• irregular structure
• implicit structure
• partial structure
• a posteriori ‘data guide’ versus a priori schema
• large data guides

Semi-structured data - properties

• It should be possible to ignore the data guide upon querying
• Data guide changes fast
• object can change type/class
• difference between data guide and data is blurred

Semi-structured data - model

• network of nodes
• object model (oid)
• query: path search in the network

OEM (Object Exchange Model)

• Graph
• Nodes: objects
  oid
  atomic or complex
  - atoms: integer, string, gif, html, …
  - value of a complex object is a set of
    object references (label, oid)
• Edges have labels
• OEM is used by a number of systems (ex. Lorel)
OEM example

Lorel query language

1. Find all places to eat Vietnamese food
   select P from RestaurantGuide.% P
   where P.category grep "Vietnamese"

2. Find the names and streets of all restaurants in Palo Alto
   select R.name, A.street from RestaurantGuide.restaurant{R}.address A
   where A.city = "Palo Alto"

Wildcards and variables
? - 0 or 1 path
+ - 1 or more paths
* - 0 or more paths
# - any path
% - 0 or more chars

Data Guides

• A structural summary over a databank that is used as a dynamic schema
• Is used in query formulation and optimization
• Is often created a posteriori
• Properties:
  – concise
  – accurate
  – convenient

Data Guides - definitions

• Label path: sequence of labels
  L1.L2. … .Ln
• Data path: alternating sequence of labels and oid:s
  L1.o1.L2.o2. … .Ln.on
• Data path d is an instance of label path l if the sequences of labels are identical in l and d.
Data Guides

• A databank can have several data guides

• Minimal data guides
  the smallest data guides

Minimal Data Guides

• Concise

• May be hard to maintain
  Example: child node for 10 with label E

Strong Data Guides

Intuitively:
"label paths that reach the same set of objects in the data model = label paths that reach the same objects in the data guide"

Strong Data Guides - definitions

Definition:
$d$ is a strong data guide for $s$ if for all label paths $l$ of $s$ it holds that $L(s,l) = L(d,l)$

There is a 1-1-mapping between target sets in the data model and nodes in a strong data guide.
Data Guides - example

Strong Data Guides - algorithm
Implementation:
- Traverse data model depth-first.
- Each time you find a new target set for label path \( l \), create a new object in the data guide.
If the target set is already represented in the data guide, do not create a new object, but link to the existing object.

Strong Data Guides - use
- Easier to maintain
- Used as path index for query optimization

Semi-structured data - exercises

Exercise 1
• Represent the relations below using the OEM data model.

<table>
<thead>
<tr>
<th>r_id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>Hamlet</td>
</tr>
<tr>
<td>r2</td>
<td>Normandie</td>
</tr>
<tr>
<td>r3</td>
<td>McDonald's</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>Linkoping</td>
</tr>
<tr>
<td>c2</td>
<td>Norkoping</td>
</tr>
</tbody>
</table>

Exercise 2
• Using the data model from the previous question, formulate the following queries using Lorel:
  – find all the restaurants that are located in Linkoping
  – find the address (city and street) of the "Hamlet" restaurant
  – list the restaurants by city (equivalent of GROUP BY)
Exercise 3

• Write 4 simple queries in Lorel that illustrate the use of coercion in the following types of comparison:
  – string type against integer type;
  – value against atomic object
  – value against complex object
  – value against set of objects

Explain how coercion works in each case

Exercise 4

• Draw the strong Data Guide for the restaurant guide data model below.