DIADEM | domain-centric intelligent automated data extraction methodology

European Research Council

erc

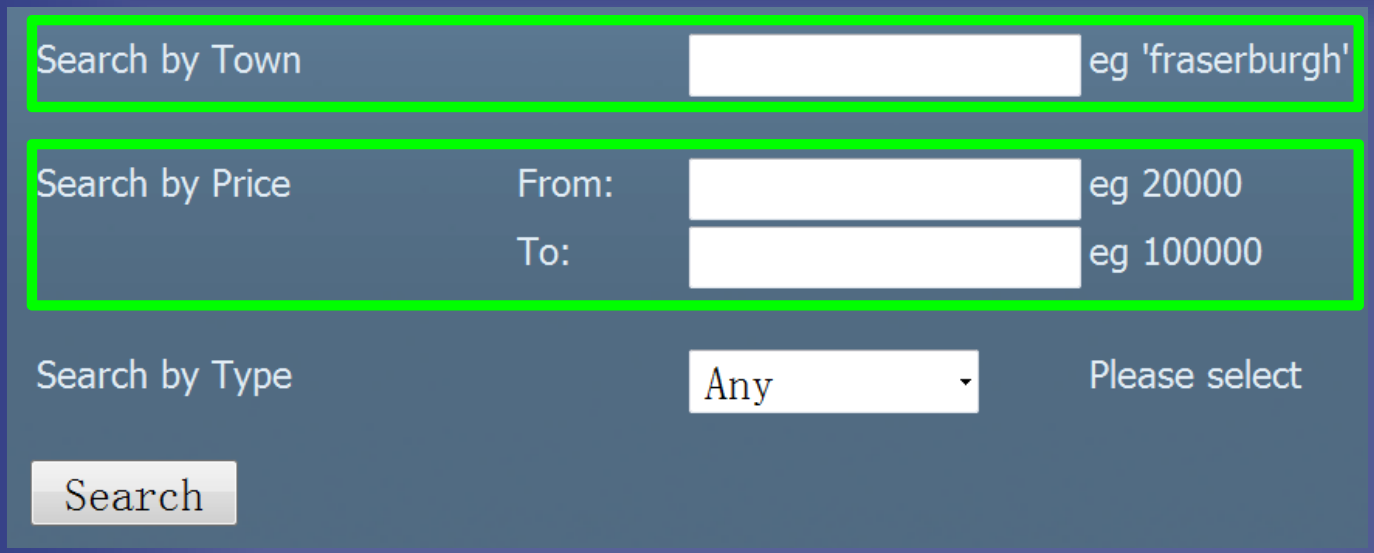
OPAL – Real Understanding of Real Estate Forms

Xiaonan Guo
Oxford University Computing Laboratory, DIADEM group

Diversity in Web Form Design

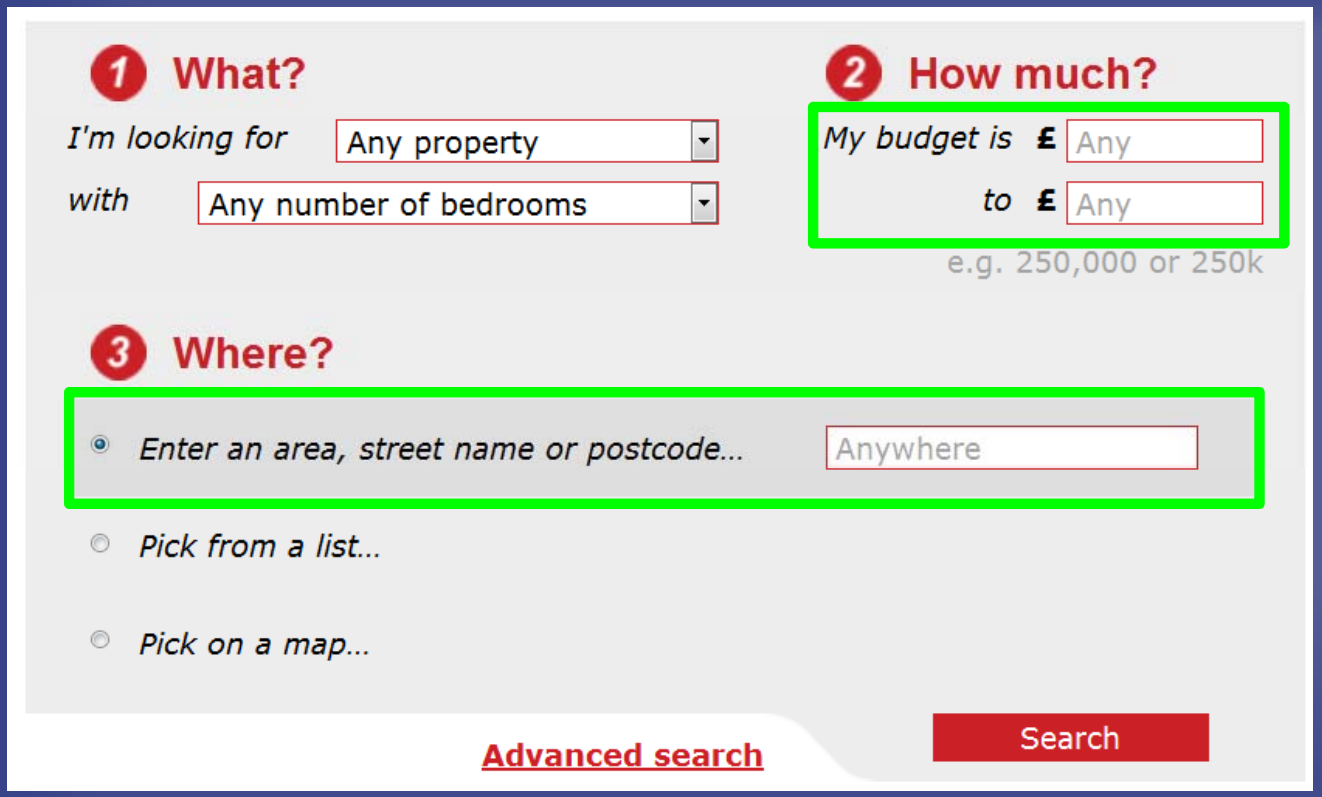
DIADEM

Diversity in Web Form Design



DIADEM

Diversity in Web Form Design



Diversity in Web Form Design

Find a property to buy or rent...

To Buy: ◉ To Rent: ○

Area:
- ○ Nailsea / Backwell
- ○ Portishead / Pill
- ○ Clevedon
- ○ Yatton / Congresbury
- ○ Bristol / Weston-super-mare

Min. beds: Select

Min. price: --- Any Price ---

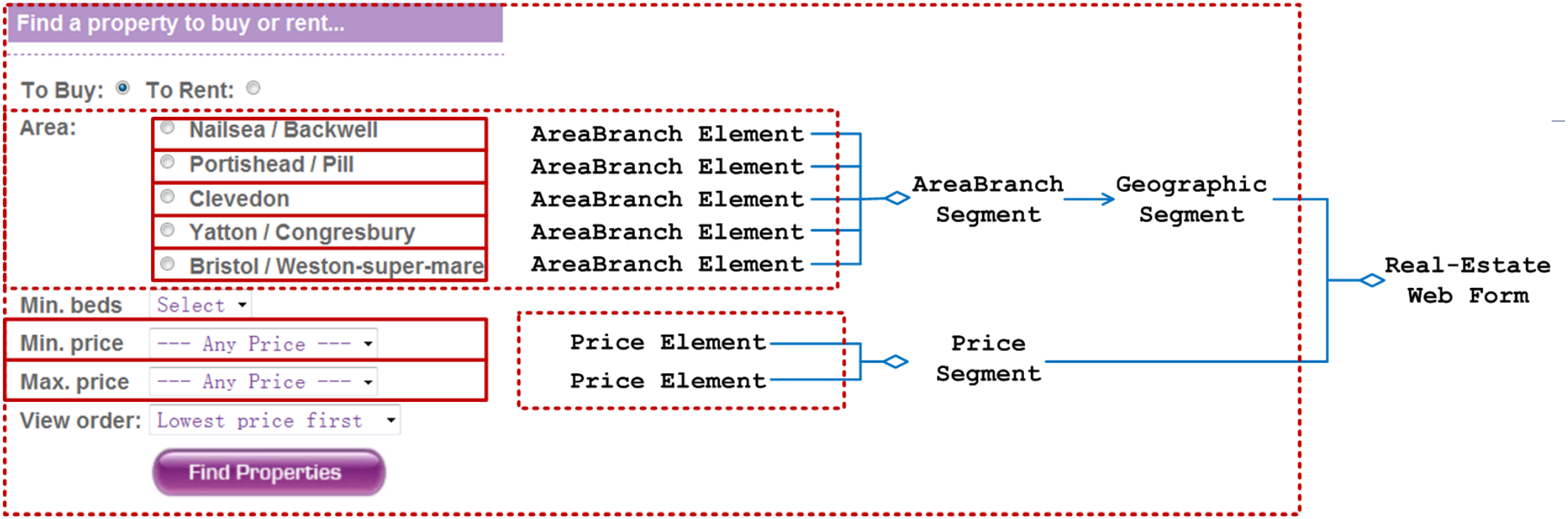
Max. price: --- Any Price ---

View order: Lowest price first

Find Properties

DIADEM

Diversity in Web Form Design



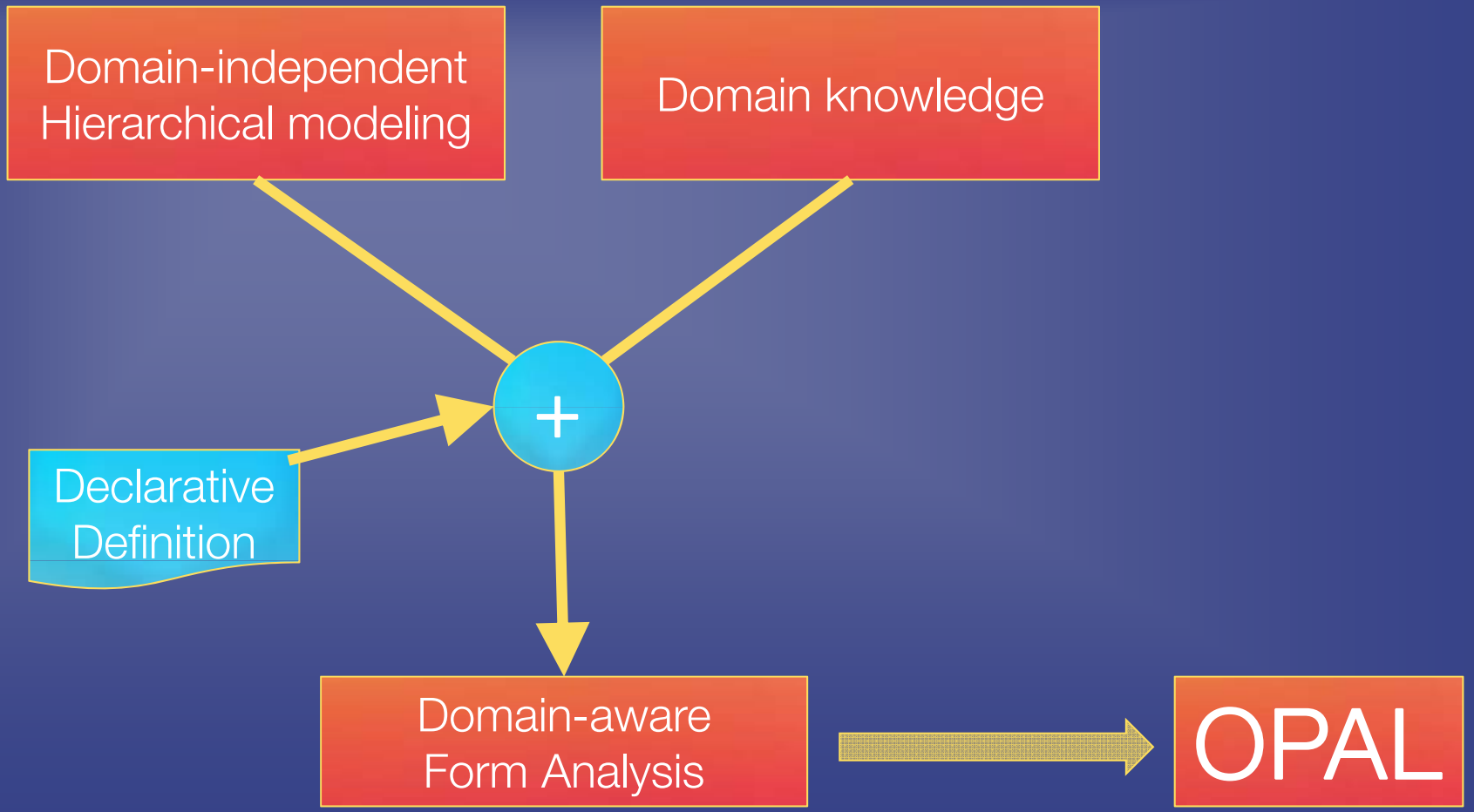
Outline

OPAL – **O**ntology-based web **P**attern **A**nalysis with **L**ogic

- Overview
- Data models and Mappings
  - Browser, Segmentation, Annotation, and Domain Model
  - Segmentation and Phenomenological Mapping
- Analysis and Evaluation
- Future Work

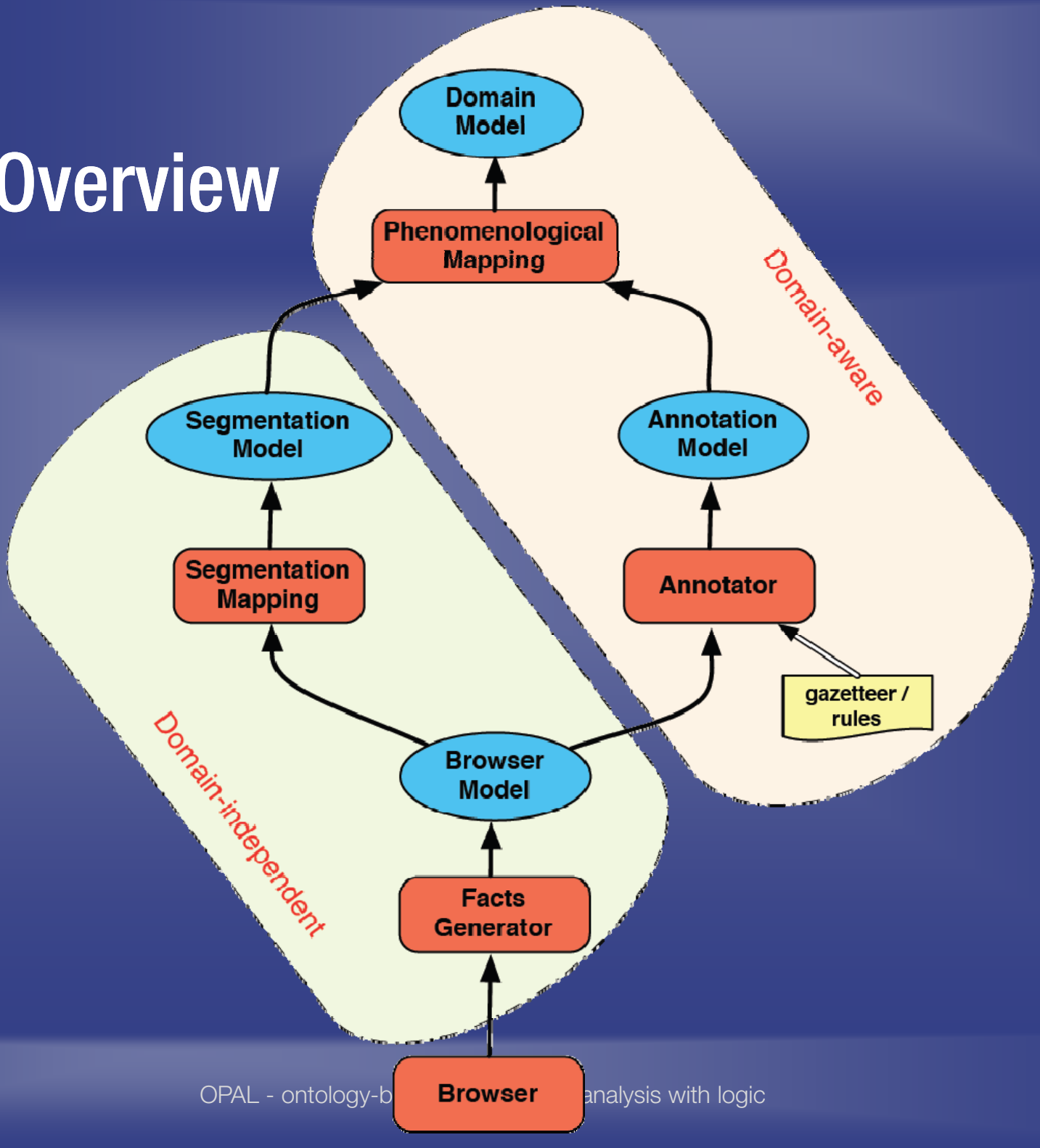

DIADEM

Overview

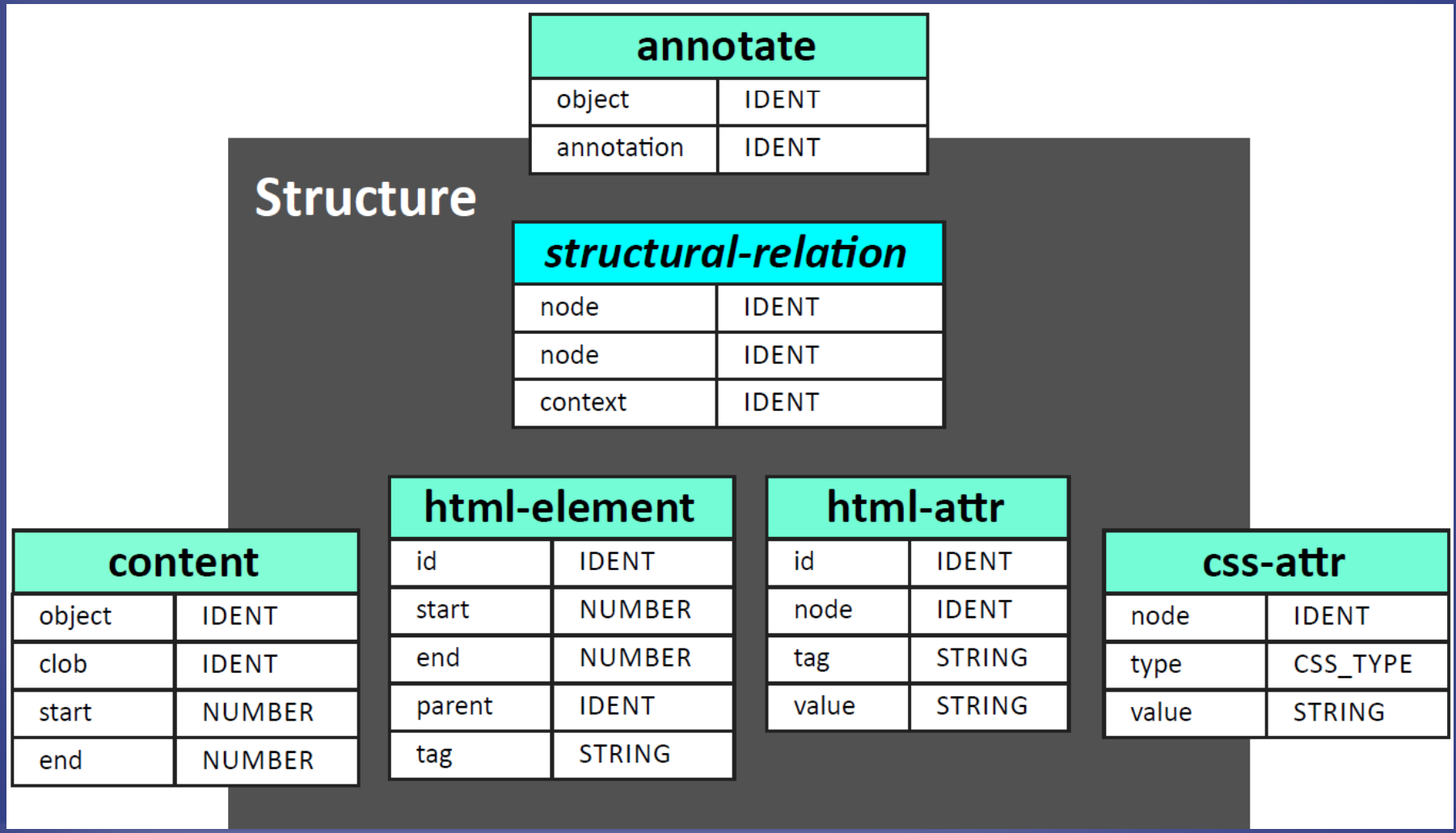

Domain-independent
Hierarchical modeling
Domain knowledge
+
Declarative
Definition
Domain-aware
Form Analysis
OPAL

DIADEM

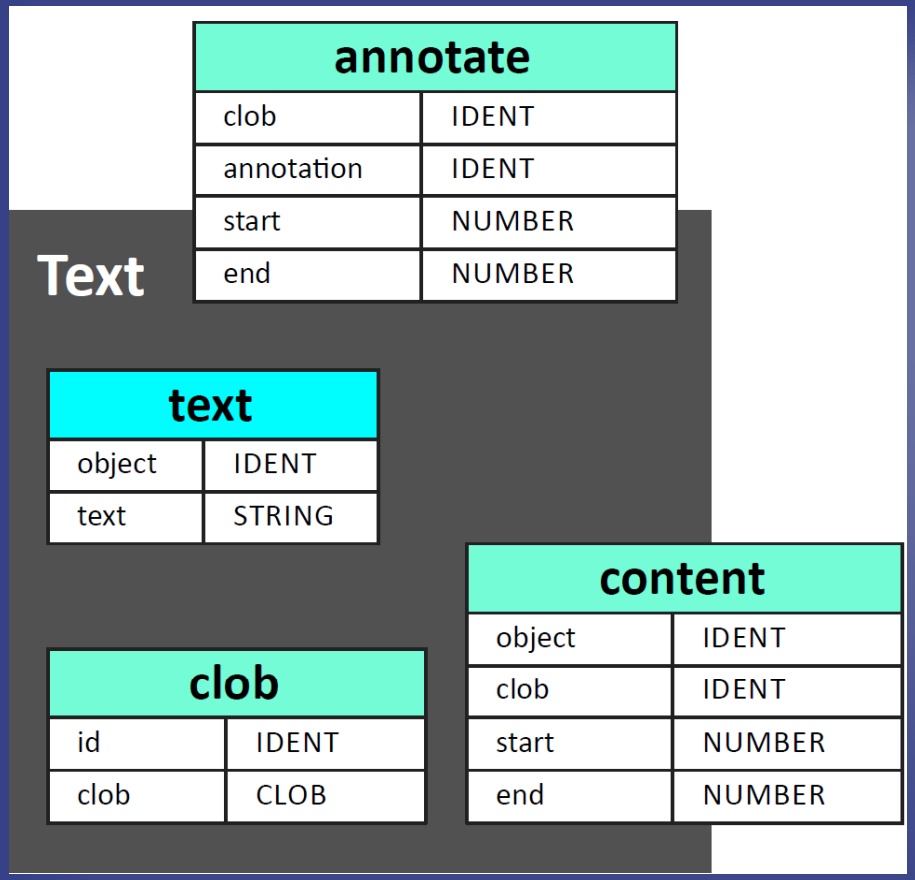
Overview

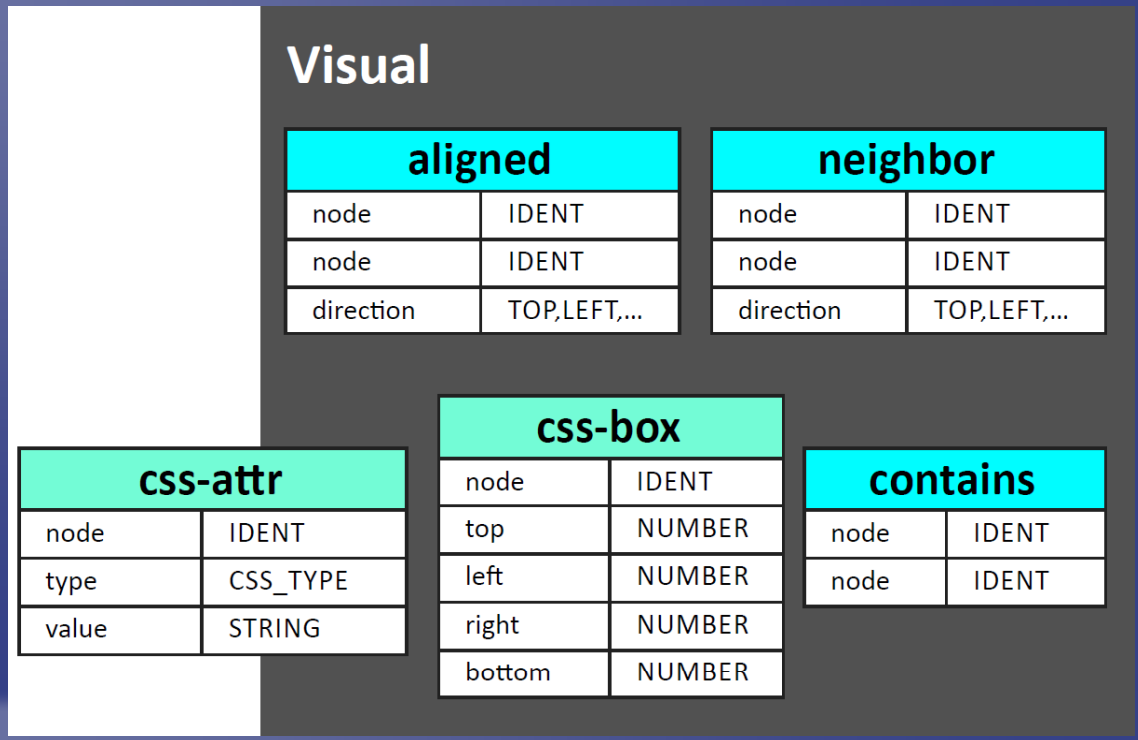


Browser Model



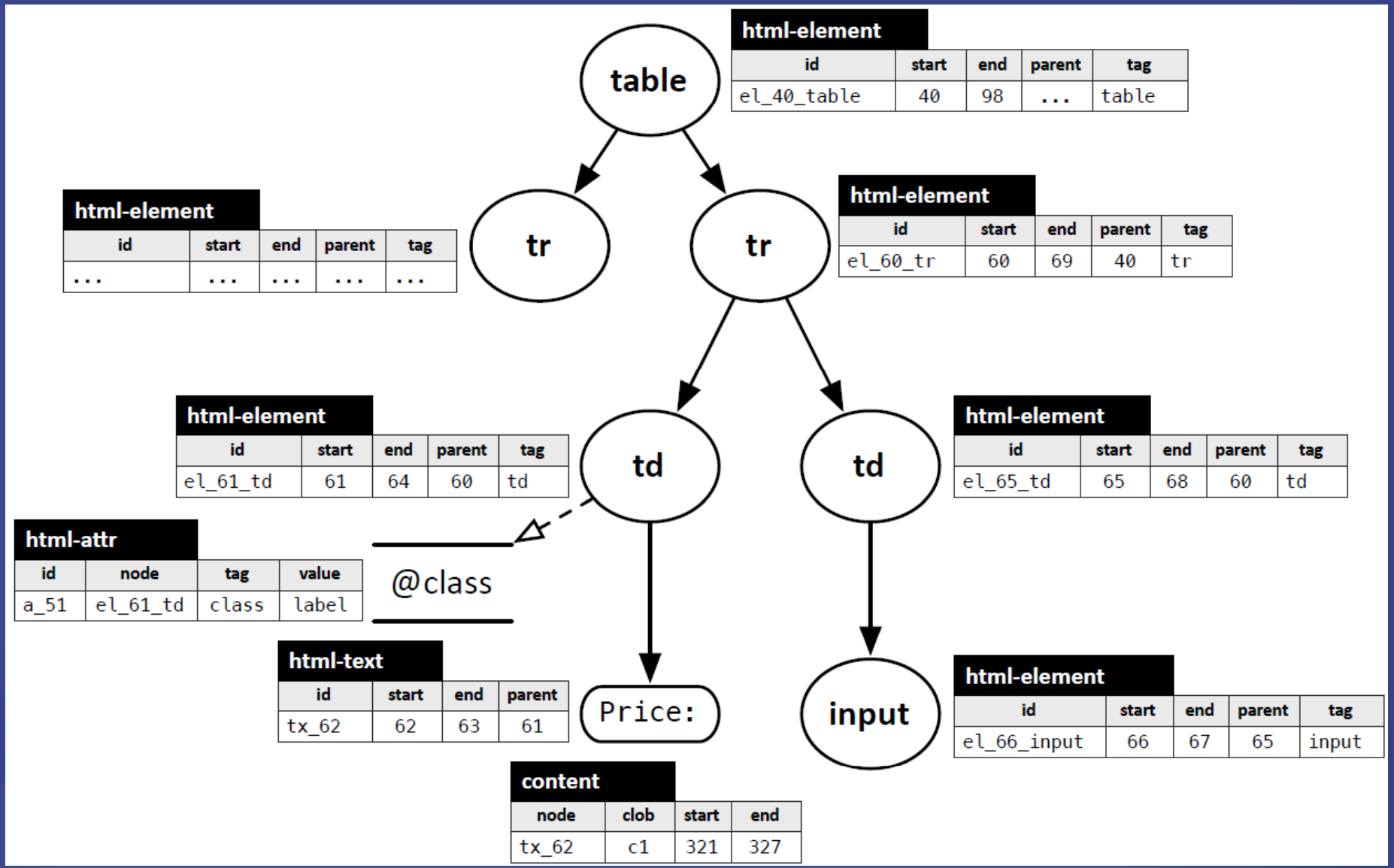
Browser Model



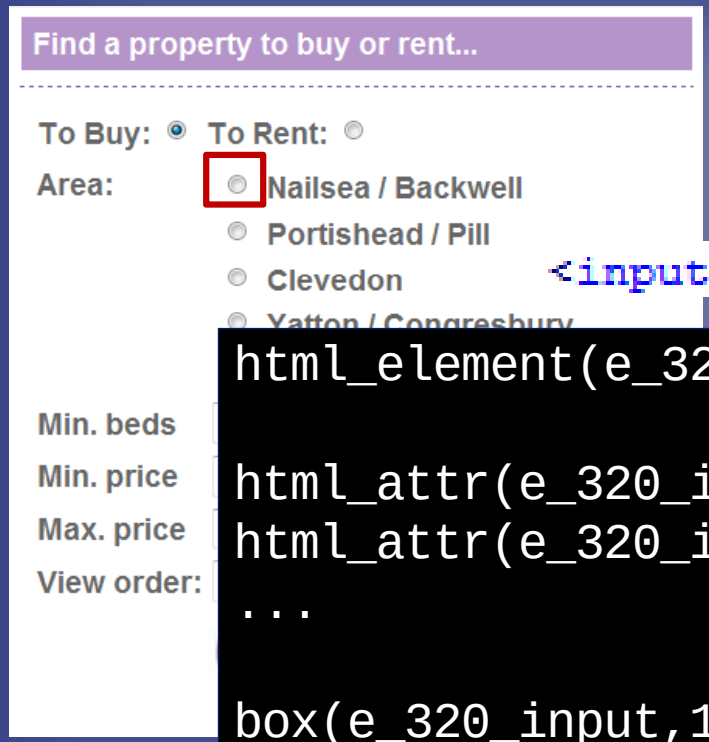


DIADEM

Browser Model



Browser Model




```
<input type="radio" value="nailsea" name="location">
```

```
html_element(e_320_input,320,321,319,input,d1).

html_attr(e_320_input_name,e_320_input,name,"location“,d1).
html_attr(e_320_input_type,e_320_input,type,"radio“,d1).
...

box(e_320_input,106,253,120,267,14,14).

css_attr(e_320_input,bottom,"auto").
css_attr(e_320_input,clear,"none").
css_attr(e_320_input,color,"rgb(0,0,0)").
...
```

Segmentation Model

- grouping of related form elements, e.g. fields, labels
- achieved via Segmentation Mapping, which
  - groups form fields
  - assigns labels to fields and groups

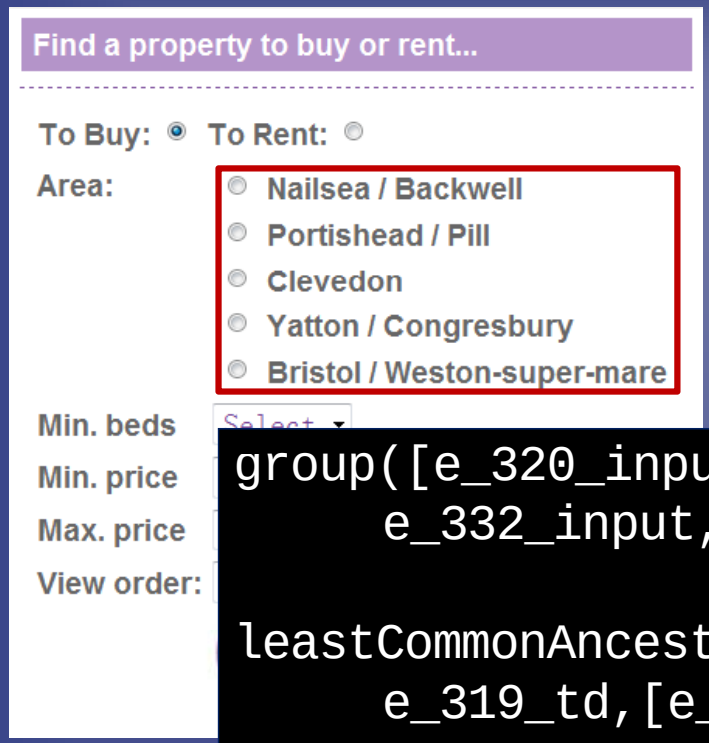
Segmentation Model – groups

- Form elements are grouped if
  - they occur in sequence
  - they have similarities in attribute values or appearances
  - their least common ancestor contains no other elements

```
group(Es) :- similarFieldSequence(Es),
     leastCommonAncestor(A,Es), not hasAdditionalField(A,Es).

leastCommonAncestor(A,Es) :- commonAncestor(A,Es),
     not ( child(C,A), commonAncestor(C,Es) ).

partOf(E,A) :-
     group(Es), member(E,Es), leastCommonAncestor(A,Es).
```

DIADEM

Segmentation Model – groups




```
group([e_320_input,e_326_input,
     e_332_input,e_338_input,e_344_input]).

leastCommonAncestor(
     e_319_td,[e_320_input,e_326_input,...,e_344_input]).

partOf(e_320_input,e_319_td).
partOf(e_326_input,e_319_td).
partOf(e_332_input,e_319_td).
partOf(e_338_input,e_319_td).
partOf(e_344_input,e_319_td).
```

Segmentation Model – labels

- Texts are assigned to fields and groups using
  - Field: HTML <label>, Greatest unique ancestor
  - Segment: Text-field alignment in groups
  - Page: Visual alignment

DIADEM

Segmentation Model – labels

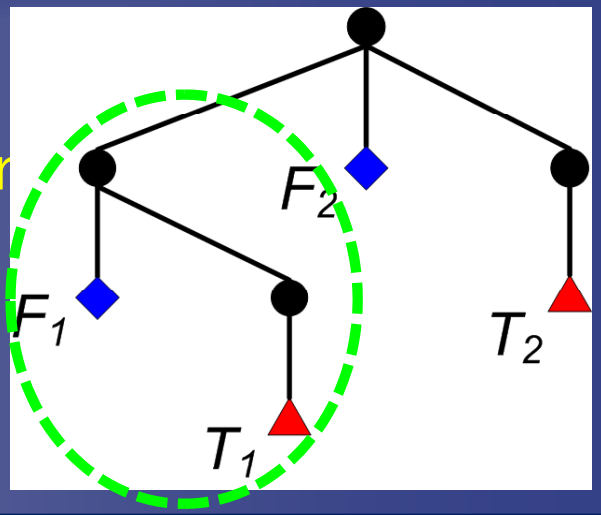
- Texts are assigned to fields and groups using
  - Field: HTML <label>, Greatest unique ancestor
  - Segment: Text-field alignment in groups
  - Page: Visual alignment

```
hasBasicLabel(E,L,T) :-
     html_element(E,_,_,_,input,_),html_attr(_,E,id,ID,_),
     html_element(N,_,_,_,label,_),html_attr(_,N,for,ID,_),
     child(L,N), html_text(L,T,_).
```

```
<input id="srchBuy" class="radio" type
<label for="srchBuy">Buy</label>
```

Segmentation Model – labels

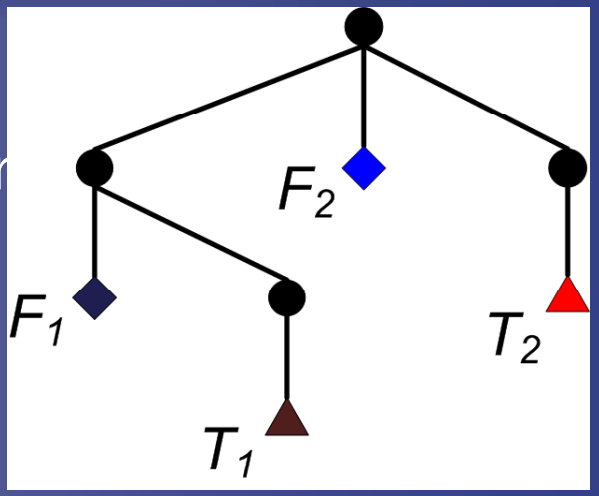
- Texts are assigned to fields and groups using
  - Field: HTML <label>, Greatest unique ancestor
  - Segment: Text-field alignment in groups
  - Page: Visual alignment




```
greatestUniqueAncestor(A,E) :- uniqueDescendant(E,A),
     not ( parent(P,A), uniqueDescendant(E,P) ).

hasBasicLabel(E,L,T) :-
     group(E), greatestUniqueAncestor(A,E),
     descendant(L,A), html_text(L,T,_).
```

Segmentation Model – labels

- Texts are assigned to fields and groups using
  - Field: HTML <label>, Greatest unique ancestor
  - Segment: Text-field alignment in groups
  - Page: Visual alignment




```
hasLabel(E,L,T) :-
     partOf(E,G), leastCommonAncestor(G,Es), group(Es),
     hasNoLabel(Es), textLists(LLs,G),
     sameLength(Es,LLs),
     labelOneToOne(E,Ls,Es,LLs),
     member(L,Ls),
     html_text(L,T,_).
```

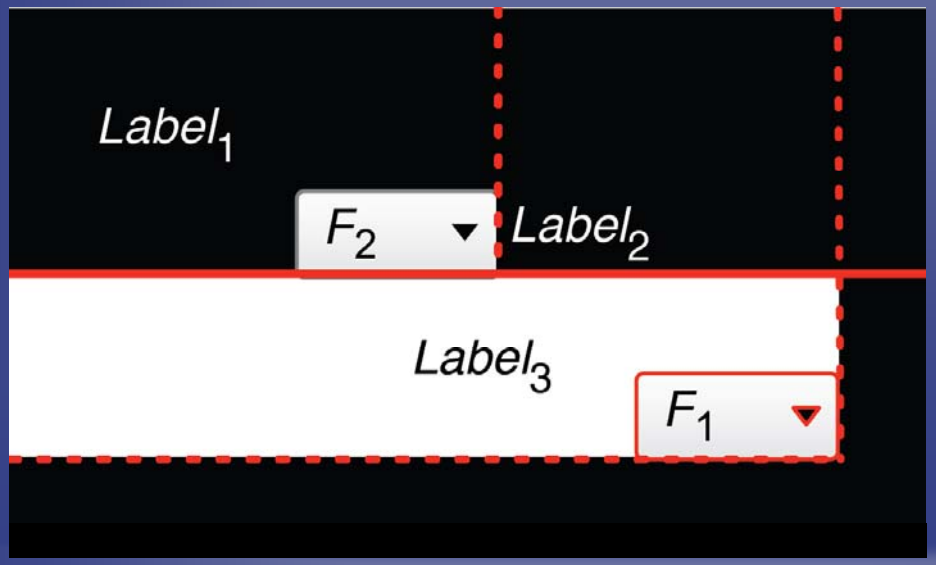
Segmentation Model – labels

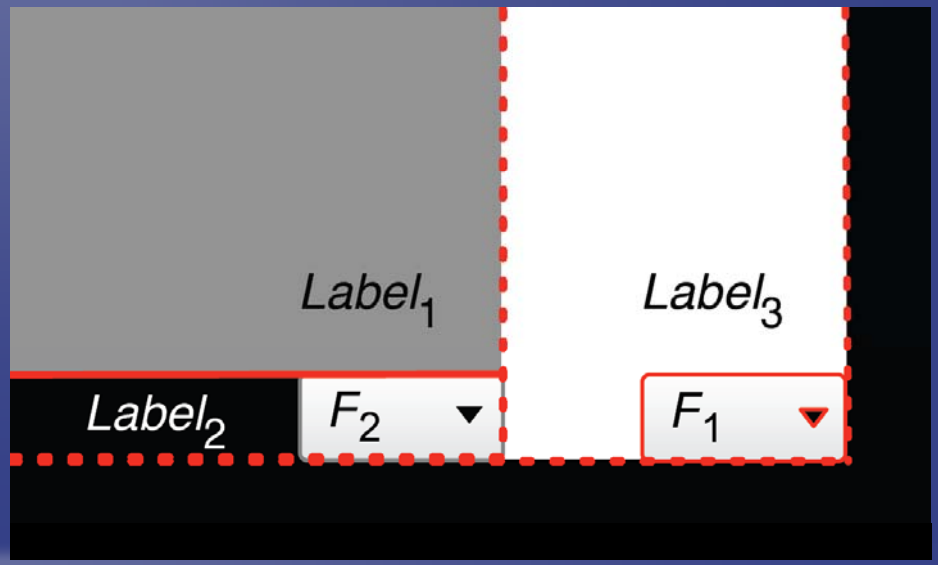
- Texts are assigned to fields and groups using
  - Field: HTML <label>, Greatest unique ancestor
  - Segment: Text-field alignment in groups
  - Page: Visual alignment

```
hasLabel(E,L,LText) :-
     hasTextBox(F,B),
     descendantTextOf(L,B),
     html_text(L,_,_,_,_),
     node_text(L,LText,true)
```

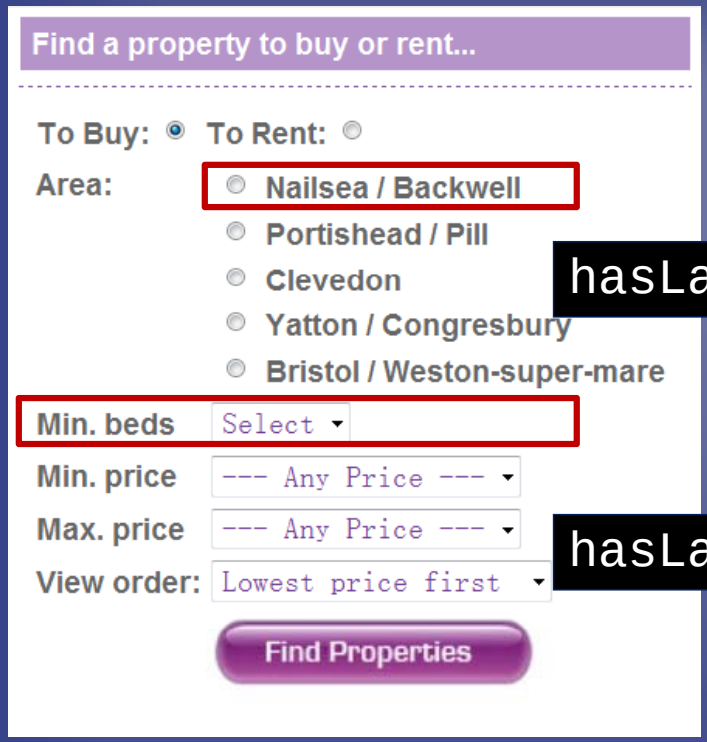
Segmentation Model – labels

- Texts are assigned to fields and groups using
  - Field: HTML <label>, Greatest unique ancestor
  - Segment: Text-field alignment in groups
  - Page: Visual alignment





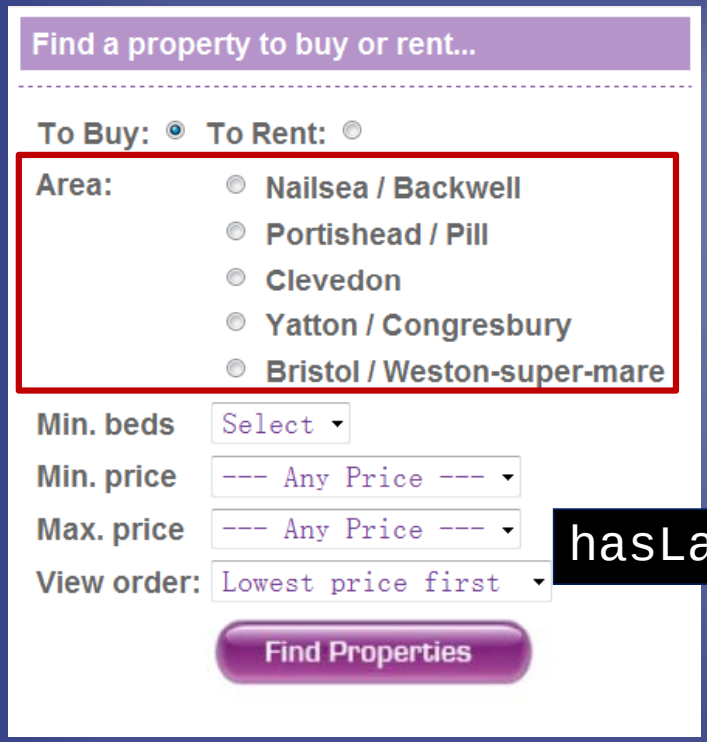
Segmentation Model – labels




```
hasLabel(e_320_input,t_322,"Nailsea / Backwell").
```

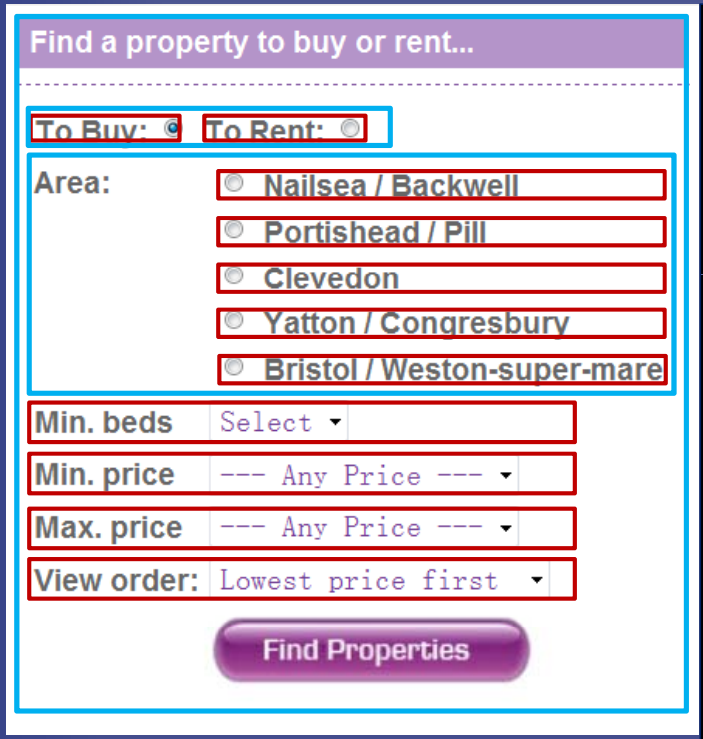
```
hasLabel(e_358_select,t_354,"Min. beds").
```

Segmentation Model – labels




```
hasLabel(e_319_td,t_316,"Area: ").
```

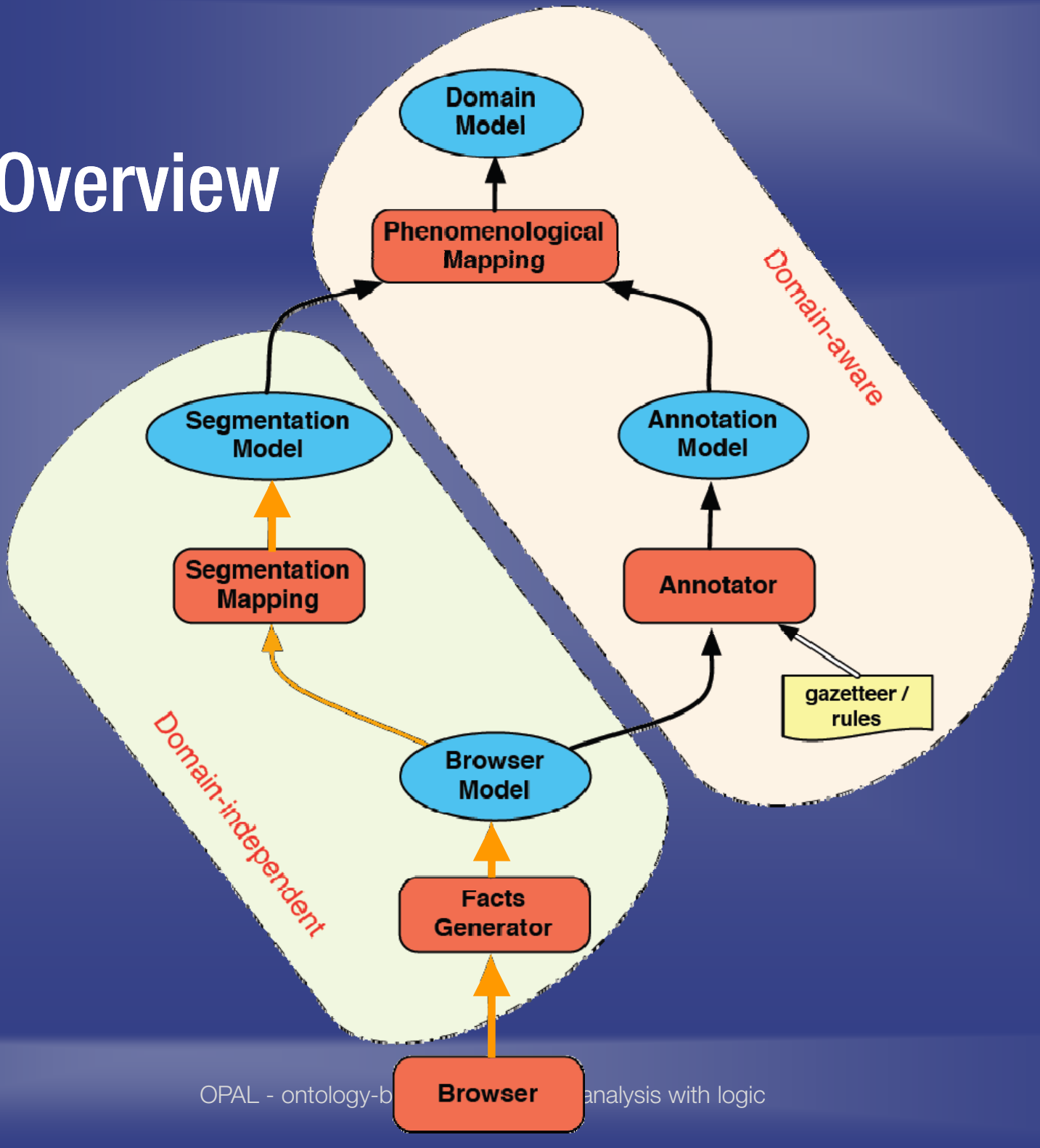
Segmentation Model – labels




```
hasLabel (e_304_input,t_302,"To Buy:").
hasLabel (e_308_input,t_306,"To Rent:").

hasLabel(e_320_input,t_322,"Nailsea / Backwell").
hasLabel(e_326_input,t_328,"Portishead / Pill").
hasLabel(e_338_input,t_340,"Yatton / Congresbury").
hasLabel(e_332_input,t_334,"Clevedon").
hasLabel(e_344_input,t_346,"Bristol / Weston-super-mare").

hasBasicLabel(e_358_select,t_354,"Min Beds").
hasBasicLabel(e_515_select,t_400,"Min Price").
hasBasicLabel(e_705_select,t_594,"Max Price").
hasBasicLabel(e_788_select,t_784,"View Order").

hasLabel(e_319_td,t_316,"Area").
hasLabel(e_297_tbody,t_290,"Find a property to buy or rent…").
```

DIADEM

Overview

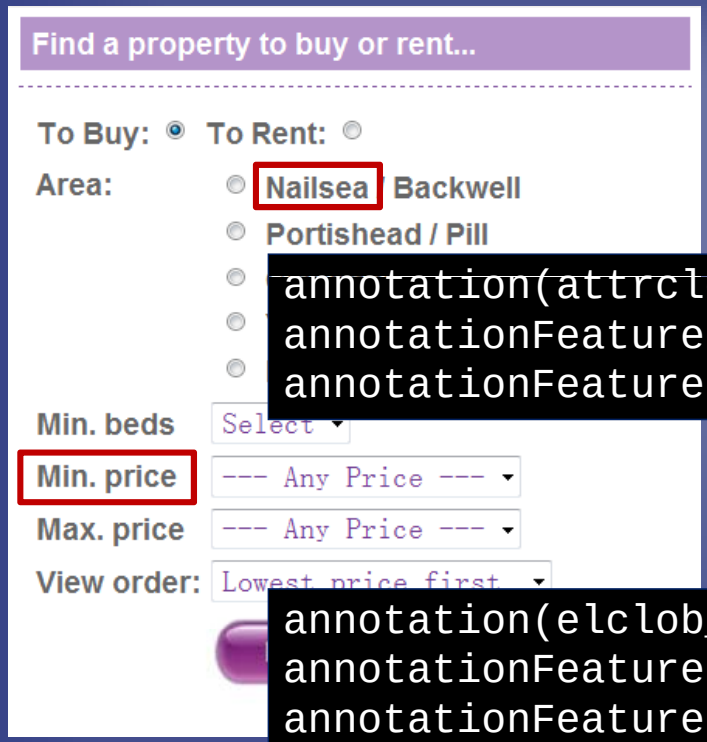


Annotation Model

- Obtained from Browser model
- Relying on domain-specific knowledge, represents
  - linguistic annotations
  - machine learning based classifications

DIADEM

Annotation Model



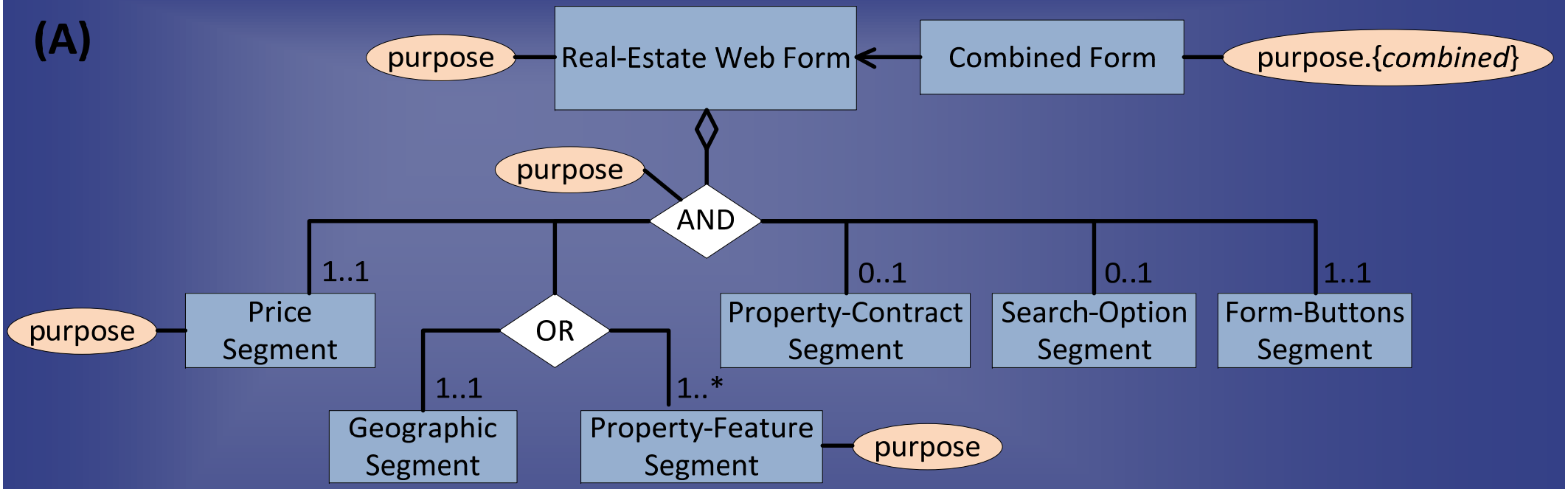

```
annotation(attrclob_d1_5560,attrclob_d1,80,87,"Nailsea").
annotationFeature(attrclob_d1_5560,"majorType","location").
annotationFeature(attrclob_d1_5560,"minorType","district_county_etc").
```

```
annotation(elclob_d1_1707,elclob_d1,2028,2038,"Min. price").
annotationFeature(elclob_d1_1707,"modifier","min").
annotationFeature(elclob_d1_1707,"minorType","price").
annotationFeature(elclob_d1_1707,"majorType","reform.label").
```

Domain Model

- Describes conceptual entities on forms as in domain ontology
- Achieved via phenomenological mapping, which
  - correlates labels with annotations
  - classifies form elements

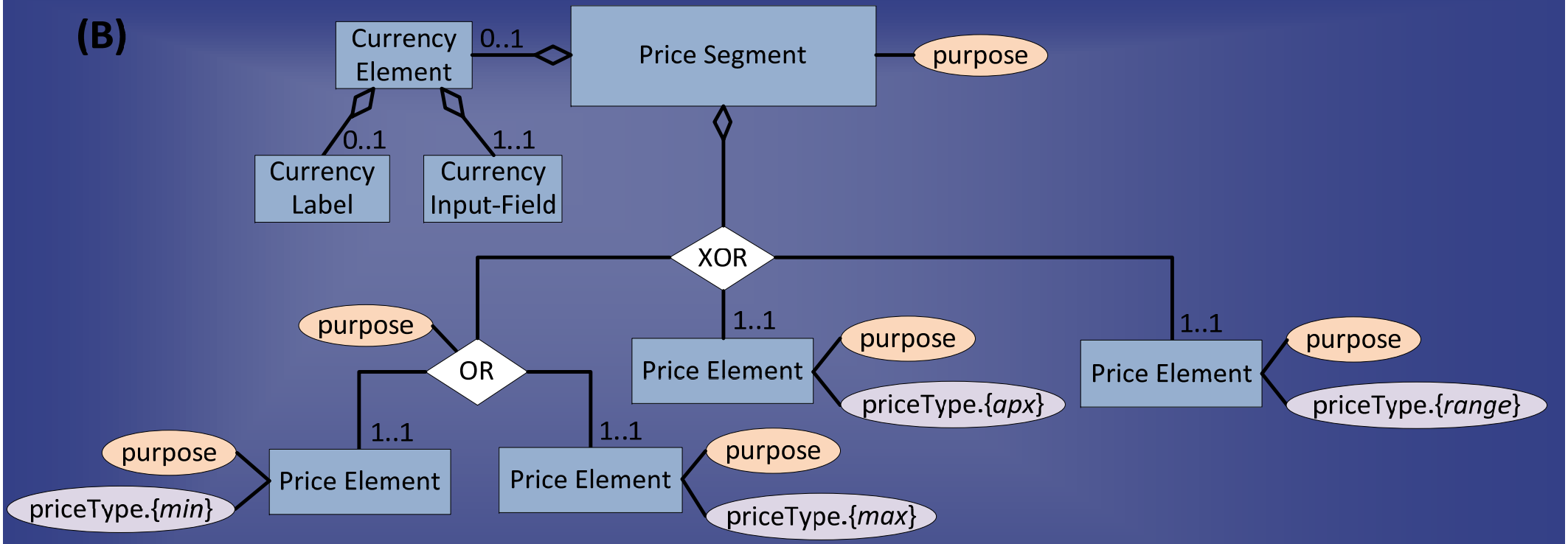
Domain Model – Ontology



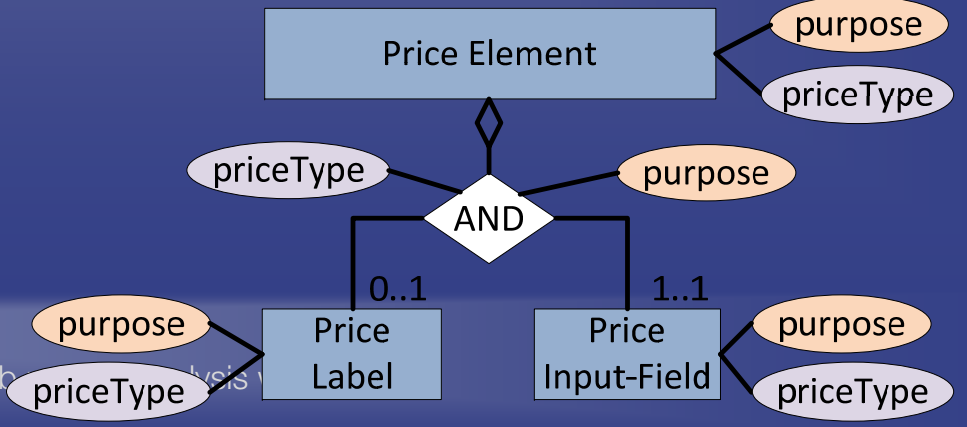
purpose={buy, rent, combined}

Domain Model – Ontology

(B)



priceType={min, max, approximate, range}



Domain Model – Classification

- Form elements are annotated as follows

```
hasAnnotation(E,A) :-
      hasLabel(E,_,T),
      annotation(Aid,_,_,_,T),
      annotationFeature(Aid,_,Anno).
```

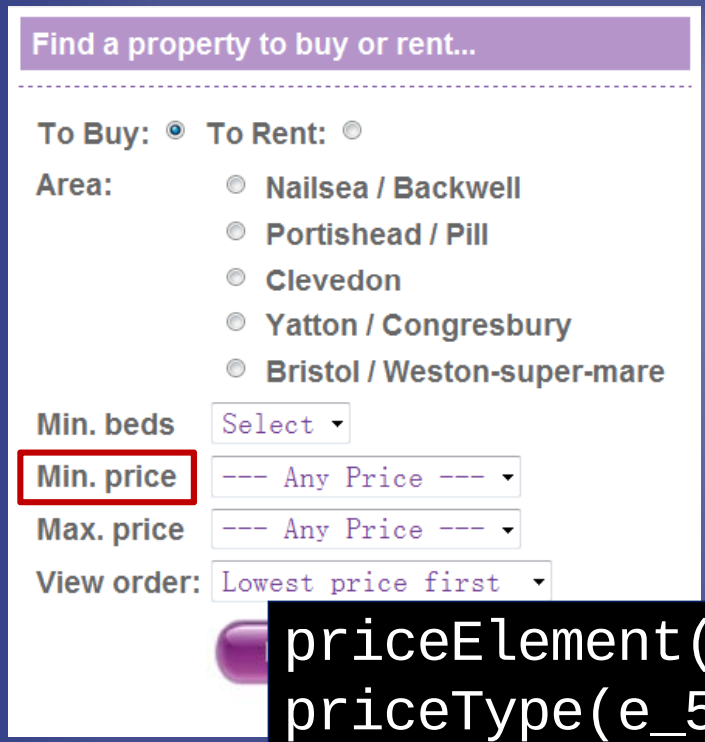
- Form elements are classified with Concept C

```
C(X) :- leafSegment(X), hasAnnotation(X,A), Clabel(A).
```

- and Facets C_F

```
C_F(X,F) :- C(X), hasAnnotation(X,F), C_FLabel(F).

C_F(X,F) :- C(X), hasValueAnnotation(X,F), C_FValue(F)
```

Domain Model – Classification



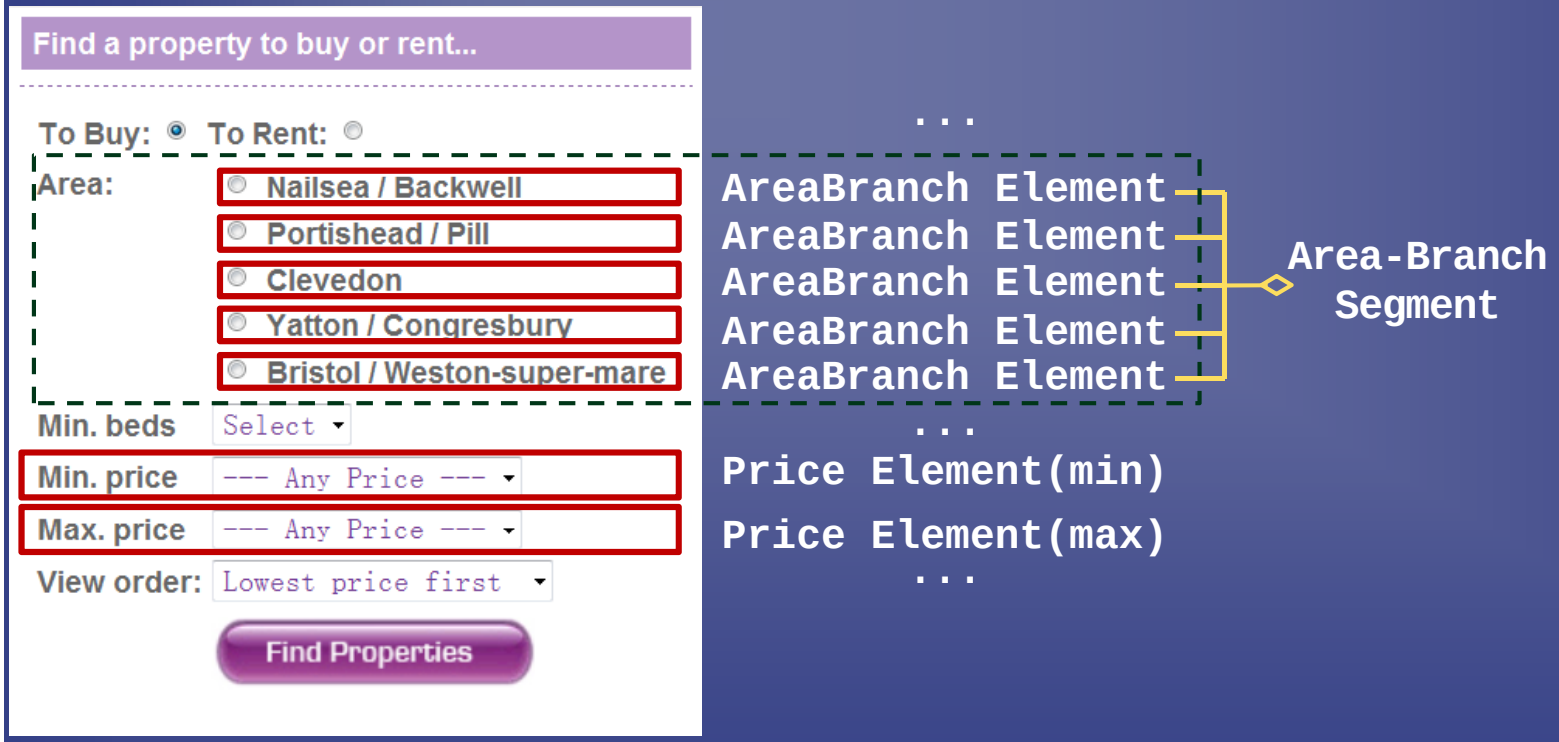

```
priceElement(e_515_select,e_286_form).
priceType(e_515_select,"min").
```

DIADEM

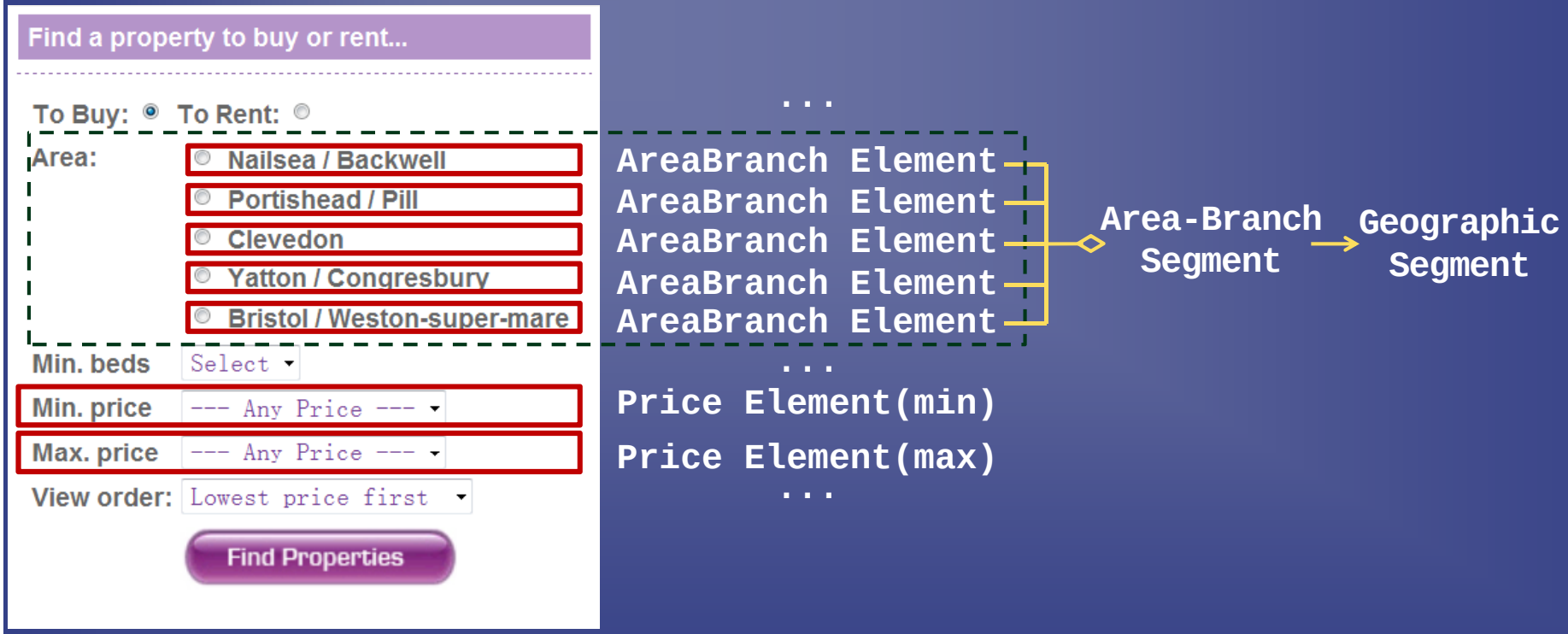
Domain Model


```
...
AreaBranch Element
AreaBranch Element
AreaBranch Element
AreaBranch Element
AreaBranch Element
...
Price Element(min)
Price Element(max)
...
```

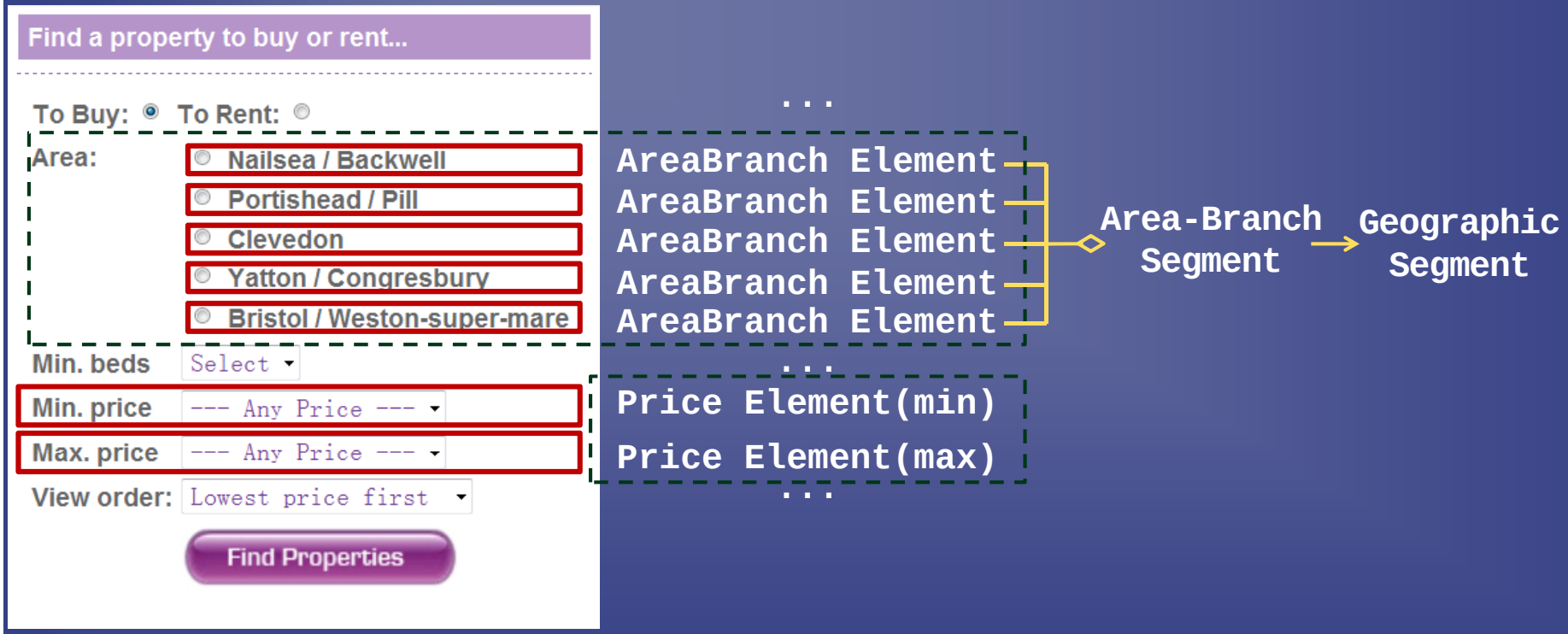
Domain Model



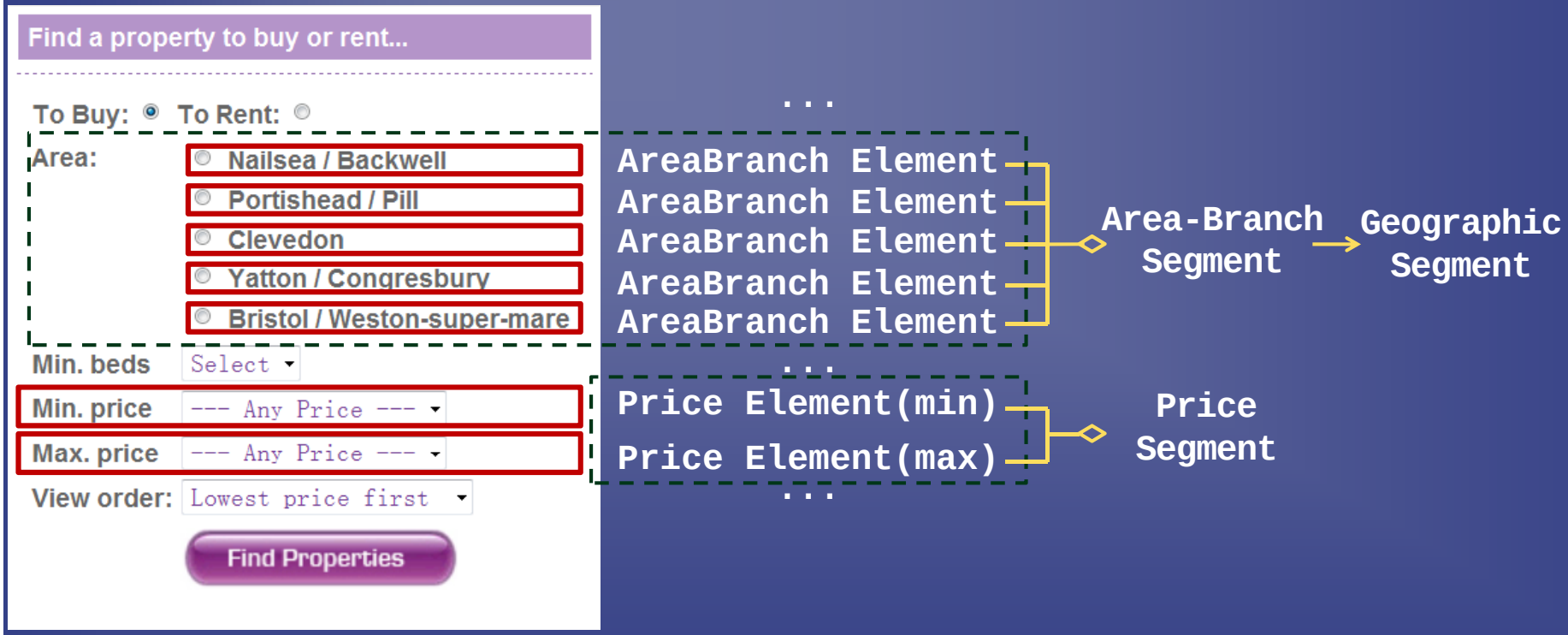
Domain Model



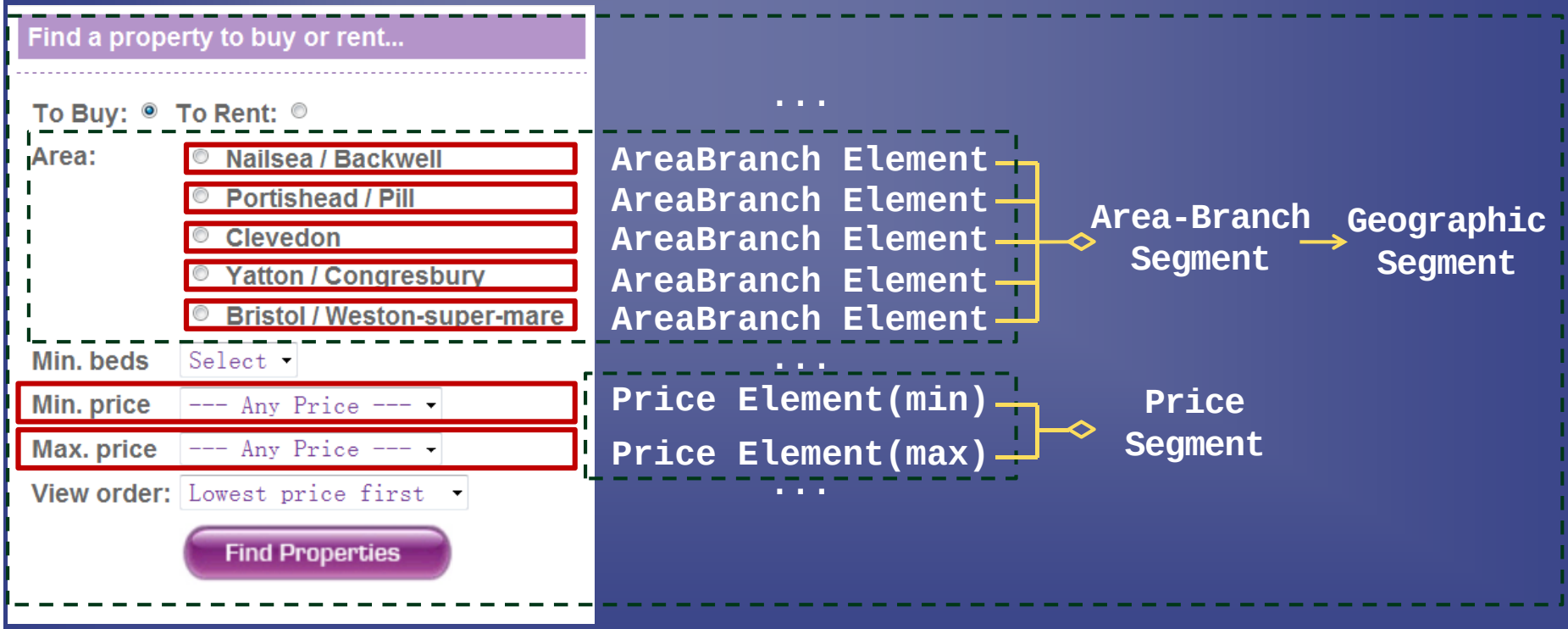
Domain Model



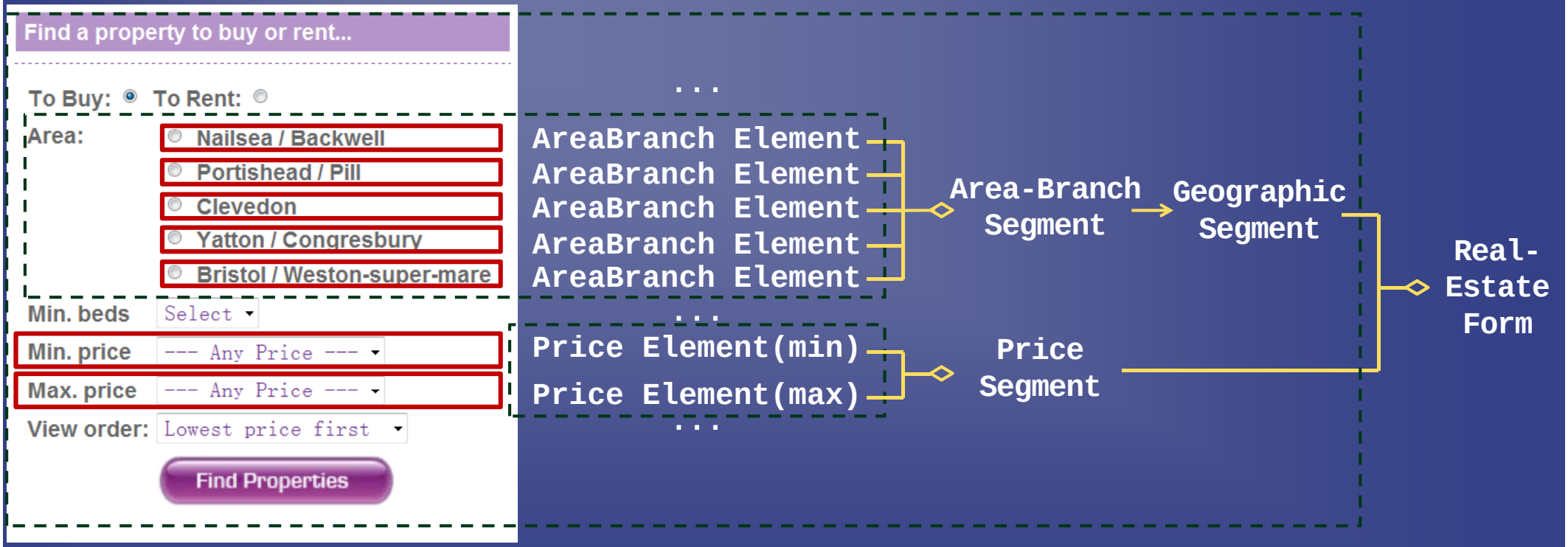
Domain Model



Domain Model

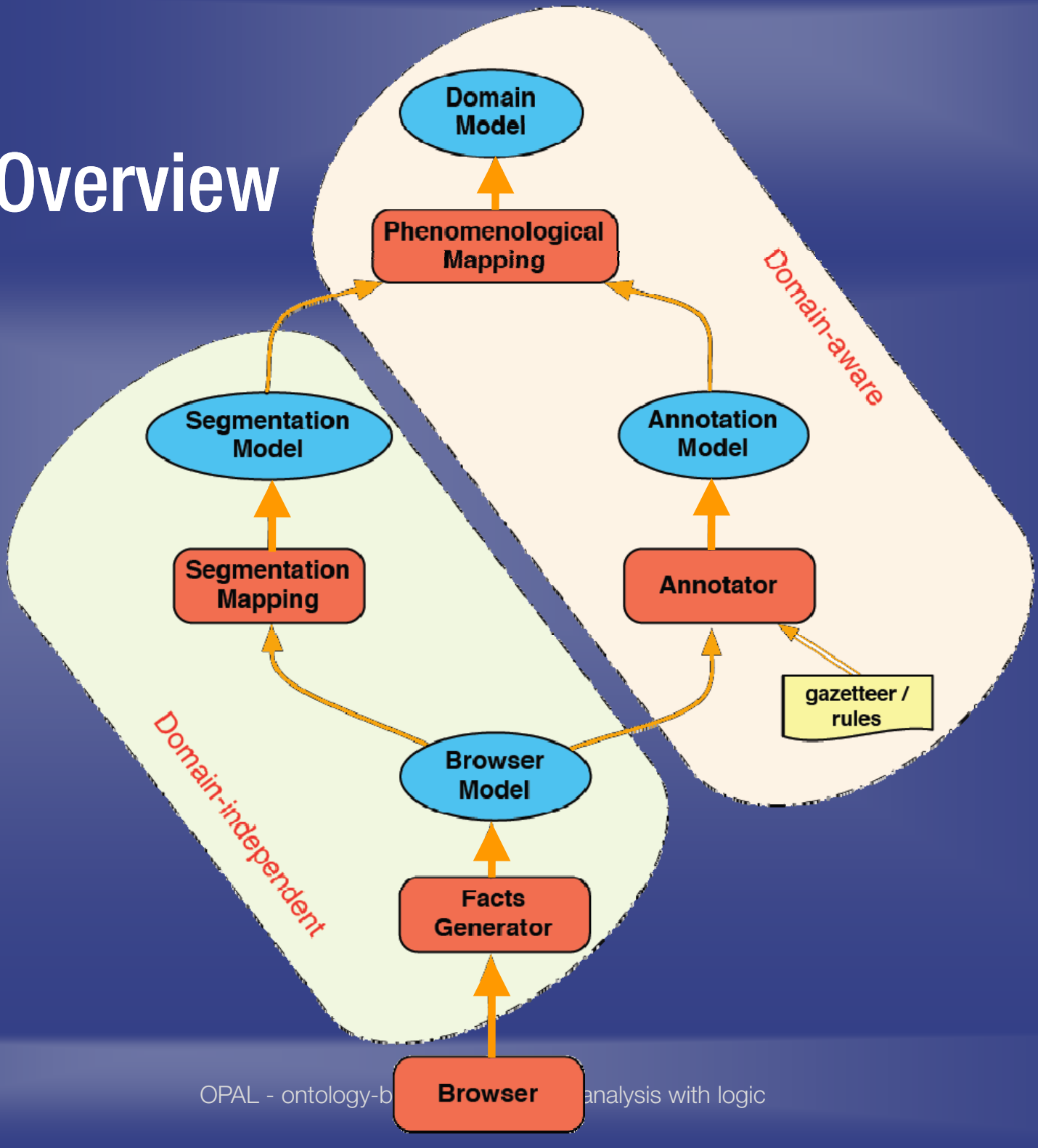


Domain Model



DIADEM

Overview



Analysis and Evaluation

- UK Real Estate Domain
  - 50 domain web forms (sampled from over 2800)
  - Tested Domain-independent and Domain-aware

DIADEM

- Real-Estate

url	form elements total	form elements found(%)	form elements labeled(%)	form segments total	form segments found
rodgersestates.com	8	100.00	100.00	1	✓
bspokeproperty.com	4	100.00	100.00	0	✓
dreweryandwheeldon.co.uk	5	100.00	100.00	0	✓
nicholasestates.co.uk	6	100.00	100.00	0	✓
wilkie.co.uk/main.htm	7	100.00	100.00	0	✓
harveyrobinson.co.uk	6	83.33	83.33	0	✓
henrygeorgeestates.co.uk	9	100.00	100.00	3	✓
dawsonsproperty.co.uk	23	100.00	100.00	6	✓
knightfrank.com	2	100.00	100.00	0	✓
all about homes.co.uk	5	100.00	100.00	0	✓
carlisleandborder.com	6	100.00	100.00	0	✓
bernadetteharris.co.uk	7	100.00	100.00	2	✓
harmony-homes.co.uk	6	100.00	100.00	0	✓
kippencampbell.co.uk	3	100.00	100.00	0	✓
jimmcmillan.co.uk	10	100.00	100.00	5	✓
huwtudor.co.uk					
tspc.co.uk					
stewartwatson.co.uk					
morganyork.co.uk					
robsoncarter.co.uk					
clearwateruk.net					
johnhoole.co.uk					
hi-m.co.uk	6	100.00	100.00	1	1 missed
qualityhomes.co.uk	13	100.00	100.00	5	✓
bychoice.co.uk	7	100.00	100.00	0	✓
rowelluk.com	9	100.00	77.78	2	1 missed
nicktart.com	17	100.00	100.00	4	✓
lawsonsestateagents.co.uk	5	100.00	100.00	1	✓
christopherbice.co.uk	7	100.00	100.00	1	✓
finders.co.uk	8	100.00	100.00	2	✓
andrewsonline.co.uk	7	100.00	100.00	0	✓
vebra.com	6	100.00	100.00	1	✓
ankerandpartners.co.uk	4	75.00	75.00	0	✓
babingtons.co.uk	5	100.00	100.00	0	✓
bairstoweves.co.uk	2	50.00	50.00	0	✓
cjhole.co.uk	7	100.00	100.00	3	✓
heritage4homes.co.uk	11	100.00	100.00	1	✓
besleyhill.co.uk	5	100.00	100.00	1	✓
countryproperty.co.uk	8	100.00	100.00	0	✓
chestertonhumberts.com	15	100.00	100.00	3	✓
edisonfordproperty.co.uk	5	100.00	100.00	0	✓
edwards-online.co.uk	7	100.00	100.00	1	✓
bruntandfussell.co.uk	5	100.00	100.00	1	✓
geoffreysmith.org	5	80.00	80.00	0	✓
sequencehome.co.uk	7	100.00	100.00	1	✓
hootons.co.uk	14	100.00	100.00	3	✓
lettingzed.co.uk	7	100.00	100.00	0	✓
houseandco.co.uk	13	92.31	84.62	6	✓
		97.61% average precision	96.68%		93.33% correct segmentation

Domain independent

Form Fields		Form Segments
found	labeled	correct segmentation
97.61%	96.68%	93.33%

DIADEM

- Real-Estate

	form fields			form groups	
url	total	found(%)	labeled(%)	total	found(%)
rodgersestates.com	8	100.00	100.00	1	v
bspokeproperty.com	4	100.00	100.00	0	v
dreweryandwheeldon.co.uk	5	100.00	100.00	0	v
nicholasestates.co.uk	6	100.00	100.00	0	v
wilkie.co.uk/main.htm	7	100.00	100.00	0	v
harveyrobinson.co.uk	6	100.00	100.00	0	v
henrygeorgeestates.co.uk	9	100.00	100.00	3	v
dawsonsproperty.co.uk	23	100.00	100.00	6	v
knightfrank.com	2	100.00	100.00	0	v
all-about-homes.co.uk	5	100.00	100.00	0	v
carlisleandborder.com	6	100.00	100.00	0	v
bernadetteharris.co.uk	7	100.00	71.43	2	v
harmony-homes.co.uk	6	100.00	100.00	0	v
kippencampbell.co.uk	3	100.00	100.00	0	v
[illegible]					
huwtudor.co.uk					
tspc.co.uk					
stewartwatson.co.uk					
morganyork.co.uk					
robsoncarter.co.uk					
clearwateruk.net					
johnhoole.co.uk	8	100.00	100.00	1	v
hi-m.co.uk	6	100.00	83.33	1	v
qualityhomes.co.uk	13	100.00	100.00	5	v
bychoice.co.uk	7	100.00	100.00	0	v
rowelluk.com	9	100.00	77.78	4	2 missed
nicktart.com	17	100.00	100.00	4	v
lawsonsestateagents.co.uk	5	100.00	100.00	1	v
christopherbice.co.uk	7	100.00	100.00	1	v
finders.co.uk	8	100.00	100.00	2	v
andrewsonline.co.uk	7	100.00	100.00	0	v
[illegible]					
bairstoweves.co.uk					
cjhole.co.uk					
heritage4homes.co.uk					
besleyhill.co.uk					
countryproperty.co.uk					
chestertonhumberts.com					
edisonfordproperty.co.uk	5	100.00	100.00	0	v
edwards-online.co.uk	7	100.00	100.00	1	v
bruntandfussell.co.uk	5	100.00	100.00	1	v
geoffreysmith.org	5	100.00	100.00	0	v
sequencehome.co.uk	7	100.00	100.00	1	v
hootons.co.uk	14	100.00	100.00	3	v
lettingzed.co.uk	7	100.00	100.00	0	v
houseandco.co.uk	12	100.00	91.67	6	v

Domain independent	Form Fields		Form Segments
	found	labeled	correct segmentation
	97.61%	96.68%	93.33%

Domain aware	Form Fields		Form Segments
	found	labeled	correct segmentation
	100.00%	97.28%	95.31%

100.00	97.28	95.31
average precision		correct segmentation

Conclusion and Future Work

- Improve structural segmentation
  - Visual segmentation and labeling
  - Ontology guided segmentation
- Accelerate domain adaption
  - Calling for machine learning for ontology creation
- Enhance ambiguity resolution
  - Necessitating probabilistic logic in the future
- Interactive form filling / probing

Thank you very much !