

OPAL – Real Understanding of Real Estate Forms

Xiaonan Guo Oxford University Computing Laboratory, DIADEM group



DIADEM Diversity in Web Form Design



UK Property, Flats & Houses for Sale browse with map		
location e.g. London, SW3, Bath, Sussex proximity This area only search		
Would you like us to email you who	· ·	
price range	property type min bedro	ooms
No Minimum No Maximum	House & Flat - Studio	•
minimum maximum		
keywords what's this?		
match any match all		







Search by Town				eg 'fraserburgh'
Search by Price	From: To:			eg 20000 eg 100000
Search by Type		Any	•	Please select
Search				







What?		2 How much?
I'm looking for	Any property	My budget is £ Any
with Any num	ber of bedrooms	to £ Any
		e.g. 250,000 or 250k
Where?Enter an area,	, street name or postcode	Anywhere
Pick from a lis	st	
O Pick on a map)	
	<u>Advanced search</u>	Search



DIADEM Diversity in Web Form Design



Find a property to buy or rent		
To Buy: Area:	To Rent: © 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



Find a prope To Buy: ◎	erty to buy or rent To Rent: ◎				
Area:	Nailsea / BackwellPortishead / PillClevedon	AreaBranch Element — AreaBranch Element — AreaBranch Element —	AreaBranch_	Geographic_	1
	Yatton / CongresburyBristol / Weston-super-mare	AreaBranch Element — AreaBranch Element —	Segment	Segment	Real-Estate
Min. beds	Select ▼		1		Web Form
Min. price	Any Price 🔻	Price Element	Price _		
Max. price	Any Price ▼	Price Element	Segment		
View order:	Lowest price first ▼	-	••		
	Find Properties				





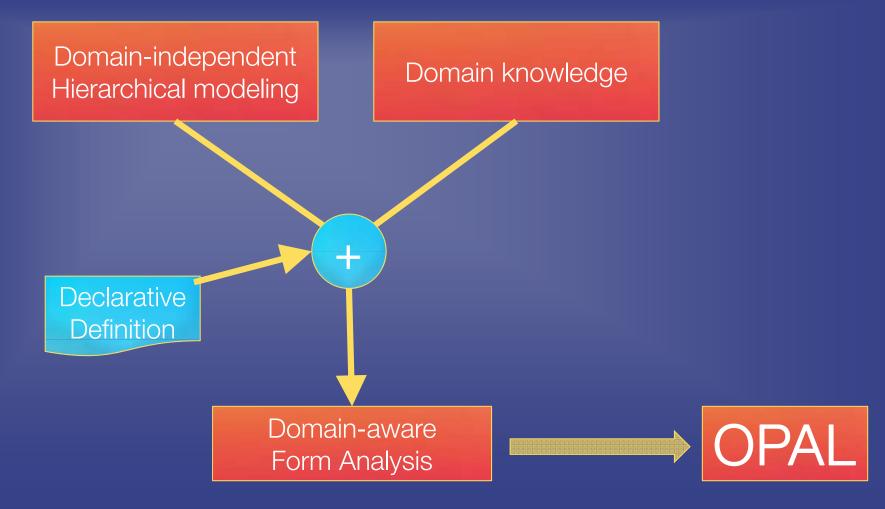
OPAL – Ontology-based web Pattern Analysis with Logic

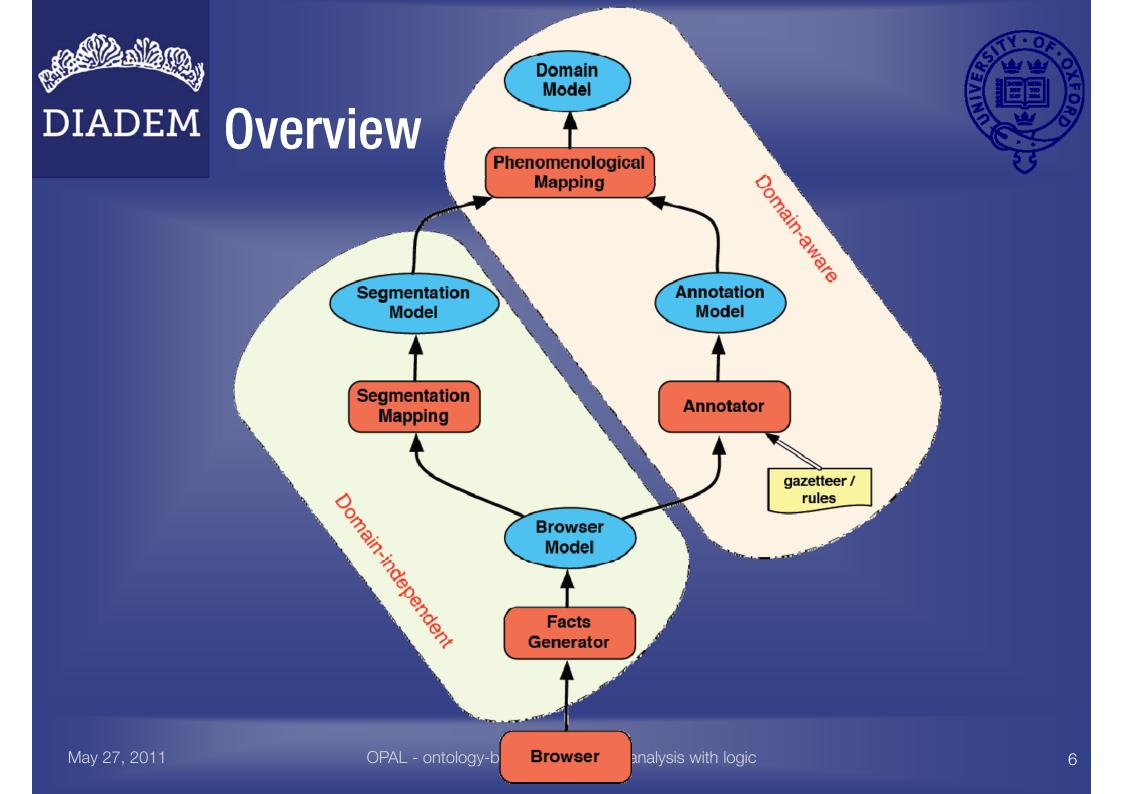
- Overview
- Data models and Mappings
 - Browser, Segmentation, Annotation, and Domain Model
 - Segmentation and Phenomenological Mapping
- Analysis and Evaluation
- **Future Work**



DIADEM Overview











annotate	
object	IDENT
annotation	IDENT

Structure

structural-relation		
node	IDENT	
node	IDENT	
context	IDENT	

content	
object	IDENT
clob	IDENT
start	NUMBER
end	NUMBER

html-element		
IDENT		
NUMBER		
NUMBER		
IDENT		
STRING		

html-attr	
id	IDENT
node	IDENT
tag	STRING
value	STRING

css-attr	
node	IDENT
type	CSS_TYPE
value	STRING





annotate	
clob	IDENT
annotation	IDENT
start	NUMBER
end	NUMBER

Text

text		
object	IDENT	
text	STRING	

clob		
id	IDENT	
clob	CLOB	

content		
object	IDENT	
clob	IDENT	
start	NUMBER	
end	NUMBER	

Visual

aligned		
node	IDENT	
node	IDENT	
direction	TOP,LEFT,	

neighbor		
node	IDENT	
node IDENT		
direction	TOP,LEFT,	

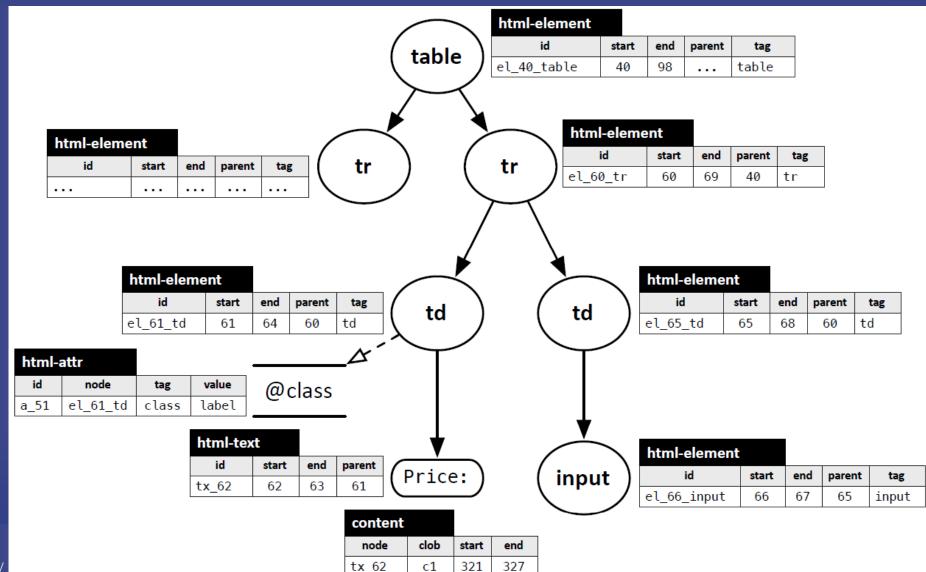
css-attr		
node	IDENT	
type	CSS_TYPE	
value	STRING	

css-box		
node	IDENT	
top	NUMBER	
left	NUMBER	
right	NUMBER	
bottom	NUMBER	

contains	
node	IDENT
node	IDENT











```
Find a property to buy or rent...
To Buy: 
To Rent:
         Nailsea / Backwell
Area:
         Portishead / Pill
                     <input type="radio" value="nailsea" name="location">
         Clevedon
        html_element(e_320_input,320,321,319,input,d1).
Min. beds
Min. price
       html_attr(e_320_input_name,e_320_input,name,"location",d1).
Max. price
        html_attr(e_320_input_type,e_320_input,type,"radio",d1).
View order:
        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)").
```



DIADEM 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



DIADEM 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



```
Find a property to buy or rent...
   To Buy: 

To Rent:
             Nailsea / Backwell
   Area:
             Portishead / Pill
             Clevedon
             Yatton / Congresbury
             Bristol / Weston-super-mare
   Min. beds
            group([e_320_input,e_326_input,
   Min. price
                  e_332_input,e_338_input,e_344_input]).
   Max. price
   View order:
            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).
May 27, 2011
```







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





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

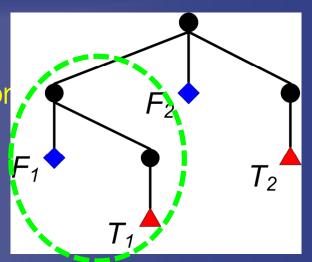
```
hasBasicLabel(E,L,T) :-
   html_element(E,_,_,_,input,_),html_attr(_,E,id,ID,_),
   html_element(N,_,_,_,label,_),html_attr(_,N,for,ID,_),
   child(L,N), html_text(L,T,_).

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





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



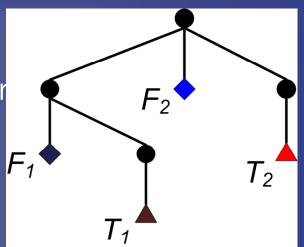
```
greatestUniqueAncestor(A,E) :- uniqueDescendant(E,A),
    not ( parent(P,A), uniqueDescendant(E,P) ).

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





- 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,_).
```







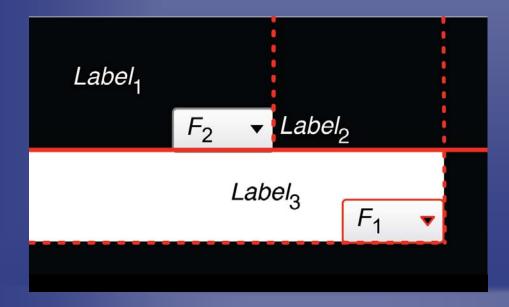
- Texts are assigned to fields and groups using
 - Field: HTML <label>, Greatest unique ancestor
 - Segment: Text-field alignment in groups
 - Page: Visual alignment

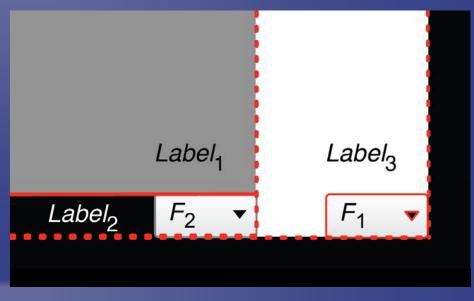
```
hasLabel(E,L,LText) :-
   hasTextBox(F,B),
   descendantTextOf(L,B),
   html_text(L,_,_,_,_),
   node_text(L,LText,true)
```





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









Find a prope	rty to buy or rent	
To Buy: Area:	To Rent: © Nailsea / Backwell Portishead / Pill	
		abel(e_320_input,t_322,"Nailsea / Backwell").
	Yatton / CongresburyBristol / Weston-super-mare	
Min. beds	Select ▼	
Min. price	Any Price ▼	
Max. price	Lowest price first hasLa	abel(e_358_select,t_354,"Min. beds").
view order.	Find Properties	





	erty to buy or rent			
To Buy:	To Rent: ©			
Area:	 Nailsea / Backwell Portishead / Pill Clevedon Yatton / Congresbury Bristol / Weston-super-mare 			
Min. beds Min. price	Select ▼ Any Price ▼			
Max. price	Lowest price first • hasLa	abel(e_319_td,t_3	316,"Area: ").	
view order.	Find Properties			





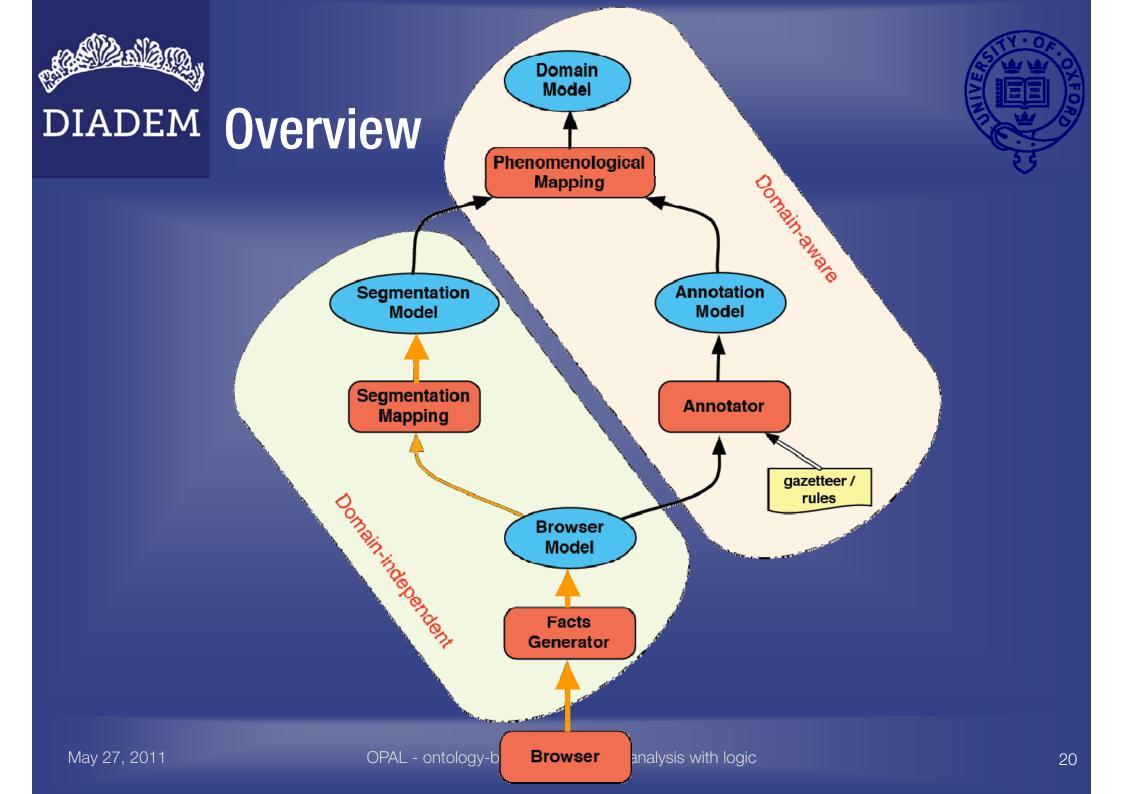
```
Find a property to buy or rent...
           To Rent: ©
To Buy: @
Area:
               Nailsea / Backwell
               Portishead / Pill
               Clevedon
               Yatton / Congresbury
            Bristol / Weston-super-mare
Min. beds
            Select ▼
Min. price
            --- Any Price --- -
           --- Any Price --- -
Max. price
View order: Lowest price first
               Find Properties
```

```
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 Annotation Model



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



DIADEM Annotation Model



```
Find a property to buy or rent...
To Buv: 
To Rent:
Area:
         Nailsea | Backwell
         Portishead / Pill
           annotation(attrclob_d1_5560,attrclob_d1,80,87 "Nailsea").
           annotationFeature(attrclob_d1_5560, "majorType", "location").
           annotationFeature(attrclob_d1_5560, "minorType", "district_county_etc"
Min. beds
        --- Any Price ---
Min. price
Max. price
         --- Anv Price ---
View order: Lowest price
           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").
```



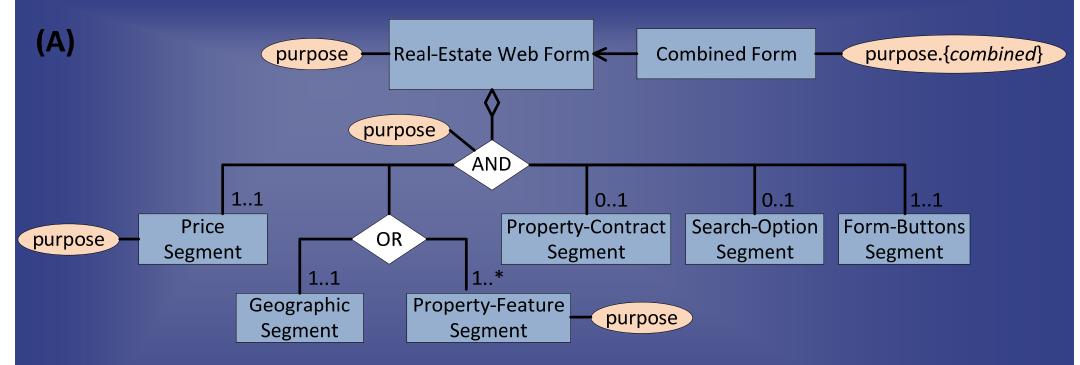


- Describes conceptual entities on forms as in domain ontology
- Achieved via phenomenological mapping, which
 - correlates labels with annotations
 - classifies form elements



DIADEM Domain Model - Ontology



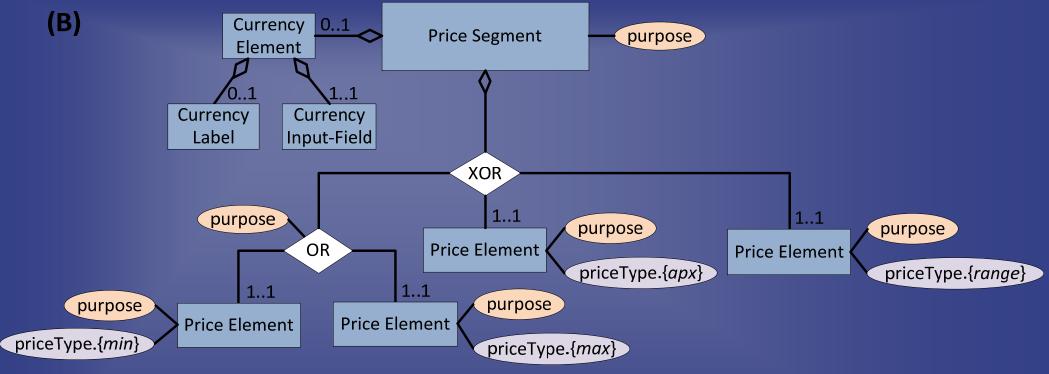


purpose={buy, rent, combined}

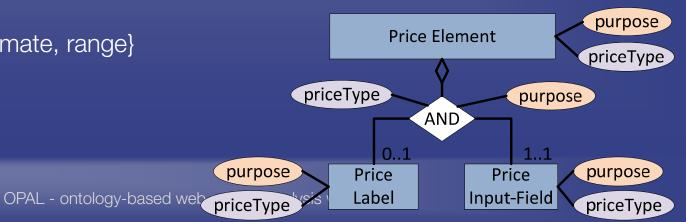


DIADEM Domain Model – Ontology





priceType={min, max, approximate, range}





DIADEM 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)
```



DIADEM Domain Model - Classification



```
Find a property to buy or rent...
To Buy: 

To Rent:
Area:
          Nailsea / Backwell
          Portishead / Pill
          Clevedon
          Yatton / Congresbury
          Bristol / Weston-super-mare
         Select ▼
Min. beds
         --- Any Price --- -
Min. price
        --- Any Price --- •
Max. price
View order: Lowest price first ▼
            priceElement(e_515_select,e_286_form).
            priceType(e_515_select, "min").
```





Find a propo	erty to buy or rent	
To Buy: ●	To Rent: ©	•••
Area:	Nailsea / Backwell	AreaBranch Element
1	Portishead / Pill	AreaBranch Element
į	Clevedon	AreaBranch Element
1	Yatton / Congresbury	AreaBranch Element
1	Bristol / Weston-super-mare	AreaBranch Element
Min. beds	Select ▼	
Min. price	Any Price ▼	Price Element(min)
Max. price	Any Price	Price Element(max)
View order:	Lowest price first •	•••
	Find Properties	





```
Find a property to buy or rent...
To Buy: 
To Rent:
                                      AreaBranch Element-
            Nailsea / Backwell
Area:
            Portishead / Pill
                                      AreaBranch Element.
                                                                     Area-Branch
            Clevedon
                                      AreaBranch Element-
                                                                        Segment
             Yatton / Congresbury
                                      AreaBranch Element-
            Bristol / Weston-super-mare
                                      AreaBranch Element
Min. beds
                                      Price Element(min)
         --- Any Price --- •
Min. price
Max. price --- Any Price --- -
                                      Price Element(max)
View order: Lowest price first ▼
             Find Properties
```





```
Find a property to buy or rent...
To Buy: 
To Rent:
                                     AreaBranch Element-
            Nailsea / Backwell
Area:
            Portishead / Pill
                                     AreaBranch Element.
                                                                    Area-Branch Geographic
                                     AreaBranch Element-
            Clevedon
                                                                       Segment
                                                                                       Segment
            Yatton / Congresbury
                                     AreaBranch Element-
            Bristol / Weston-super-mare
                                     AreaBranch Element
Min. beds
                                     Price Element(min)
         --- Any Price --- •
Min. price
Max. price --- Any Price --- ▼
                                     Price Element(max)
View order: Lowest price first ▼
            Find Properties
```





Find a prop	erty to buy or rent			
To Buy: ◎ Area:	To Rent: Nailsea / Backwell Portishead / Pill Clevedon Yatton / Congresbury	AreaBranch Element— AreaBranch Element— AreaBranch Element— AreaBranch Element—	Area-Branch Geograpl Segment Segmen	
Min. beds	■ Bristol / Weston-super-mare Select ▼	AreaBranch Element		
Min. price	Any Price ▼	Price Element(min)		
Max. price	Any Price 🕶	Price Element(max)		
View order:	Lowest price first ▼			
	Find Properties			





Find a property to buy or rent		
To Buy: To Rent: Area: Nailsea / Backwell Portishead / Pill Clevedon Yatton / Congresbury Bristol / Weston-super-mare	AreaBranch Element AreaBranch Element AreaBranch Element AreaBranch Element AreaBranch Element AreaBranch Element	Geographic Segment
Min. beds Select Min. price Any Price	Price Element(min) — Price	
Max. price Any Price ▼ View order: Lowest price first ▼	Price Element(max) Segment	
Find Properties		



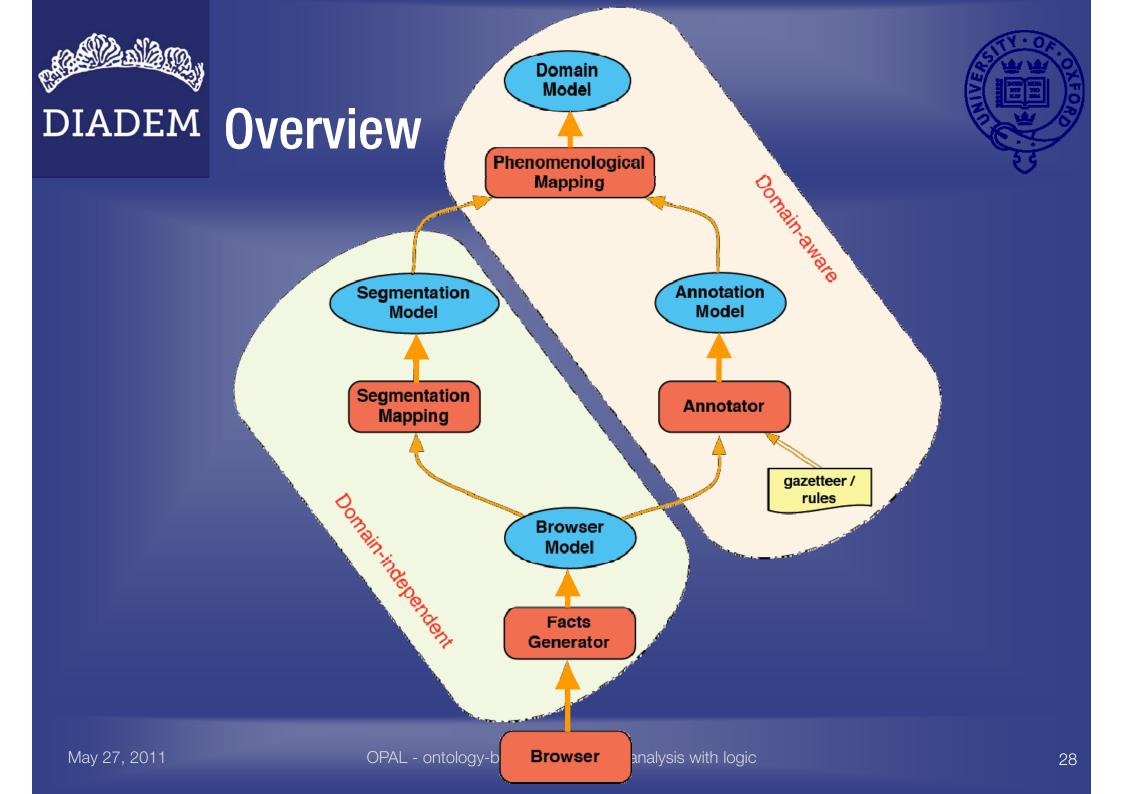


Find a property to buy or rent	
To Buy: ◎ To Rent: ◎	
Area: Nailsea / Backwell	AreaBranch Element—
Portishead / Pill	AreaBranch Element + Area-Branch Geographic
Clevedon	AreaBranch Element
Yatton / Congresbury	AreaBranch Element — Segment Segment
Bristol / Weston-super-mare	AreaBranch Element —
Min. beds Select -	
Min. price Any Price ▼	Price Element(min) — Price
Max.price Any Price →	Price Element(max) Segment
View order: Lowest price first ▼	
Find Properties	





Find a property to buy or rent		
To Buy: To Rent:		
Area: Nailsea / Backwell Portishead / Pill Clevedon Yatton / Congresbury	AreaBranch Element AreaBranch Element AreaBranch Element AreaBranch Element Segment Segment	Real-
■ Bristol / Weston-super-mare Min. beds Select ▼		→ Estate Form
Min. price Any Price ▼ Max. price Any Price ▼	Price Element(min) Price Price Element(max) Segment	
View order: Lowest price first Find Properties	•••	





DIADEM Analysis and Evaluation



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



	form elements			form segments	
url	total	found(%)	labeled(%)	total	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	✓
iimmemillan eo uk	10	100.00	100.00	5	1

Real-Estate

Domain independent

ain independent	Form Fields						
tspc.co.uk stewartwatson.co.uk morganyork.co.uk		found			labeled		
robsoncarter.co.uk clearwateruk.net johnhoole.co.uk		97.61%			96.68%		
hi-m.co.uk qualityhomes.co.uk	6 13	100.00 100.00	100.00 100.00	1 5	1 missed		
bychoice.co.uk	7	100.00	100.00	0	v		
rowelluk.com	9	100.00	77.78	2	1 missed		
nicktart.com	17	100.00	100.00	4	√,		
lawsonsestateagents.co.uk	5 7	100.00 100.00	100.00 100.00	1	√		
christopherbice.co.uk finders.co.uk	8	100.00	100.00	2	v		
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	✓		

average precision

correct segmentation

Form Segments correct segmentation 93.33%



		form fields		form	groups
url	total	found(%)	labeled(%)	total	found(%)
rodgersestates.com	8	100.00	100.00	1	٧
bspokeproperty.com	4	100.00	100.00	0	v
dreweryandwheeldon.co.uk	5	100.00	100.00	0	٧
nicholasestates.co.uk	6	100.00	100.00	0	٧
wilkie.co.uk/main.htm	7	100.00	100.00	0	٧
harveyrobinson.co.uk	6	100.00	100.00	0	٧
henrygeorgeestates.co.uk	9	100.00	100.00	3	٧
dawsonsproperty.co.uk	23	100.00	100.00	6	٧
knightfrank.com	2	100.00	100.00	0	٧
all-about-homes.co.uk	5	100.00	100.00	0	v
carlisleandborder.com	6	100.00	100.00	0	٧
bernadetteharris.co.uk	7	100.00	71.43	2	٧
harmony-homes.co.uk	6	100.00	100.00	0	٧
kippencampbell.co.uk	3	100.00	100.00	0	٧
iimmemillan eo uk	10	100.00	100.00	E	-1

- Real-Estate

Domain independent

huwtudor.co.uk		
tspc.co.uk		f,
stewartwatson.co.uk		1,
morganyork.co.uk		
robsoncarter.co.uk		Ω-
clearwateruk.net		3
johnhoole.co.uk	0	
hi-m.co.uk	6	
qualityhomes.co.uk	13	
bychoice.co.uk	7	
rowelluk.com	9	
nicktart.com	17	
lawsonsestateagents.co.uk	5	
christopherbice.co.uk	7	
finders.co.uk	8	
andrewsonline.co.uk	7	

Form Fields labeled ound 7.61% 96.68% 100.00 83.33 100.00 100.00 100.00 100.00 100.00 77.78 2 missed 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00

Form Fields

Domain aware

bairstoweves.co.uk cjhole.co.uk heritage4homes.co.uk		found		labeled 97.28%	
besleyhill.co.uk countryproperty.co.uk chestertonhumberts.com	-	L00.00	%		
edisonfordproperty.co.uk	5	100.00	100.00	0	٧
edwards-online.co.uk	7	100.00	100.00	1	٧
bruntandfussell.co.uk	5	100.00	100.00	1	٧
geoffreysmith.org	5	100.00	100.00	0	٧
sequencehome.co.uk	7	100.00	100.00	1	V
hootons.co.uk	14	100.00	100.00	3	٧
lettingzed.co.uk	7	100.00	100.00	0	٧
houseandco.co.uk	12	100.00	91.67	6	٧
		100.00	97.28	95	.31

average precision

Form Segments correct segmentation

93.33%

Form Segments

correct segmentation

95.31%

ogic

correct segmentation



DIADEM 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!