



Maria Vargas-Vera, E.Motta, J. Domingue, S. Buckingham Shum and M. Lanzoni



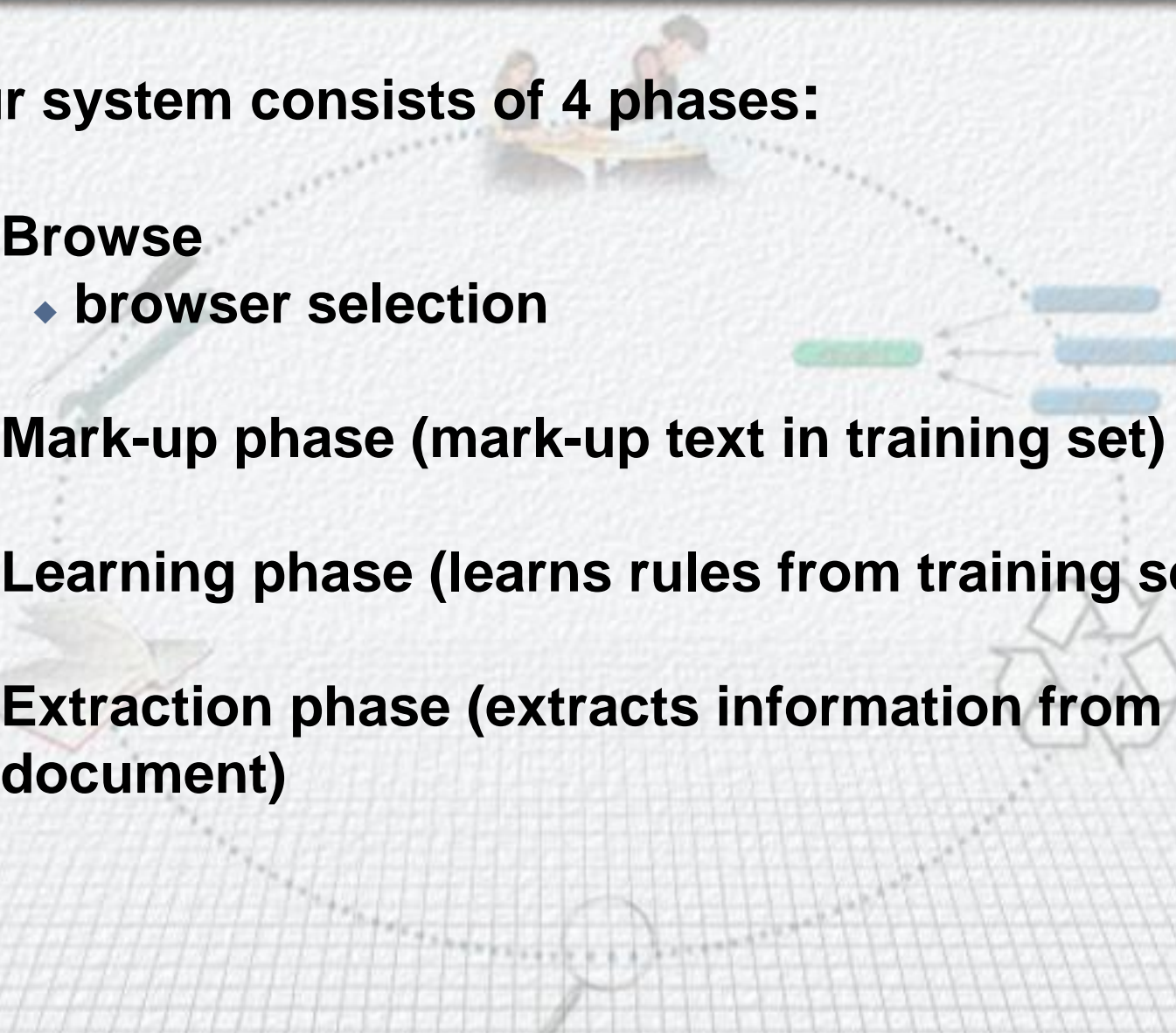
Knowledge Media Institute(KMi)

The Open University

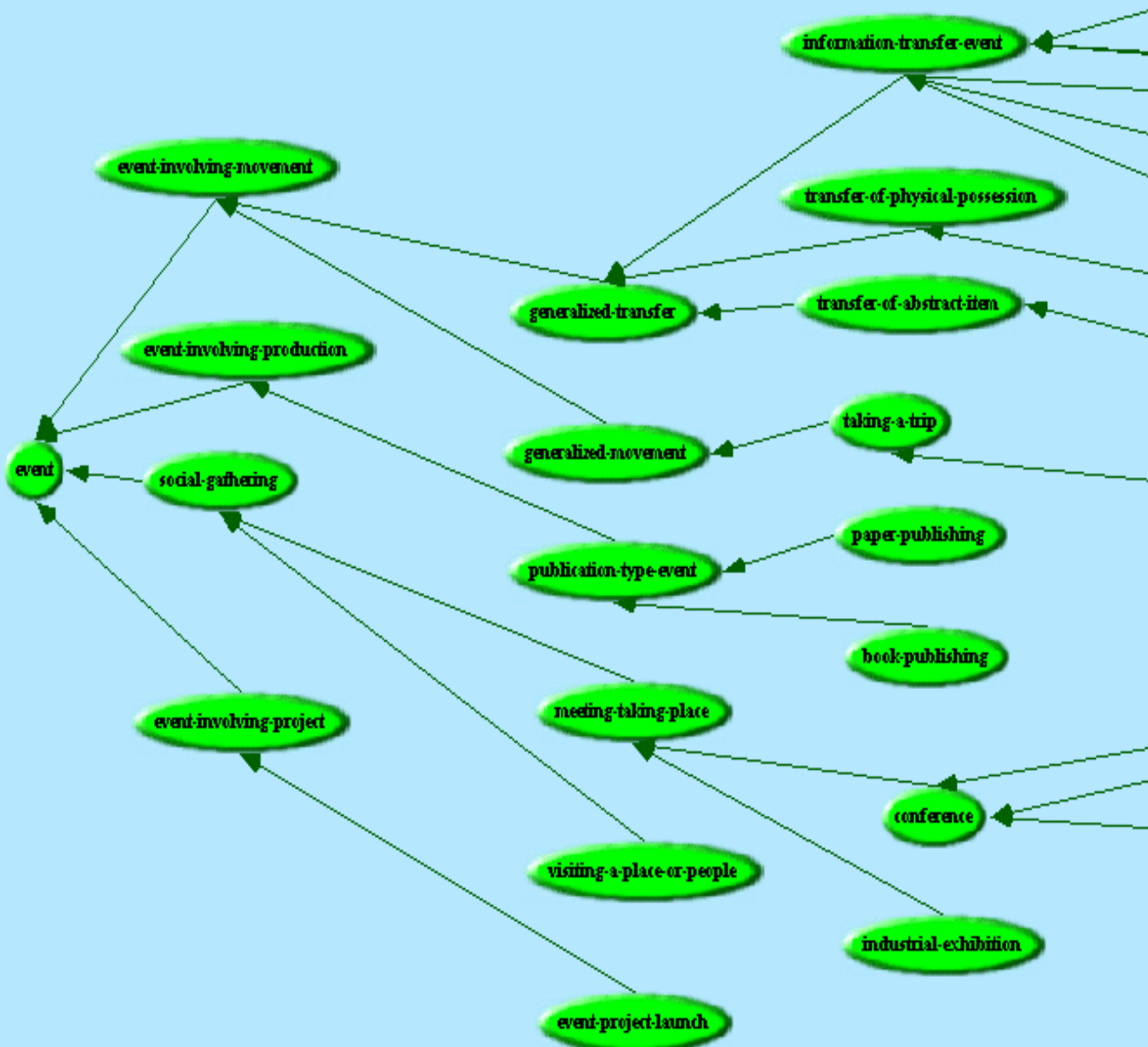
Milton Keynes, MK7 6AA

October 2001

- ◆ **Motivation**
 - ◆ Extraction of knowledge structures from web pages
 - ◆ Final goal -Ontology population
- ◆ **Approaches to semantic annotation of web pages (SAW)**
 - ◆ OntoAnnotate [Stab, et al]
 - ◆ SHOE [Hendler et al]
- ◆ **Our solution to SAW problem**
 - ◆ Ontology driven annotation
- ◆ **Work so far - we had tried with two different domains (KMi stories and Rental adverts)**
- ◆ **Conclusions and Future work**

- ◆ **Our system consists of 4 phases:**
 - ◆ **Browse**
 - ◆ **browser selection**
 - ◆ **Mark-up phase (mark-up text in training set)**
 - ◆ **Learning phase (learns rules from training set)**
 - ◆ **Extraction phase (extracts information from a document)**
- 

- ◆ **Ontology-based Mark-up**
 - ◆ **The user is presented with a set of tags (taken from ontology)**
 - ◆ **user selects slots-names for tagging.**
 - ◆ **Instances are tagged by the user**



EVENT 1:

- visiting-a-place-or-people**
- visitor** (list of person(s))
- people-or-organisation-being-visited** (list of person(s) or organisation)
- has-duration** (duration)
- start-time** (time-point)
- end-time** (time-point)
- has-location** (a place)
- other agents-involved** (list of person (s))
- main-agent** (list of person (s))

- ◆ Learning phase was Implemented using Marmot and Crystal.
- ◆ Mark-up all instances in the training set
- ◆ Marmot performs segmentation of a sentence: noun phrases, verbs and prepositional phrases.
- ◆ Example: “David Brown, the Chairman of the University for Industry Design and Implementation Advisory Group and Chairman of Motorola, visited the OU”.
- ◆ Marmot output:
 - ◆ SUBJ: DAVID BROWN %comma% THE CHAIRMAN OF THE UNIVERSITY
 - ◆ PP: FOR INDUSTRY DESIGN AND IMPLEMENTATION ADVISORY GROUP AND CHAIRMAN OF MOTOROLA
 - ◆ PUNC: %COMMA%
 - ◆ VB: VISITED
 - ◆ OBJ: THE OU

- **Crystal derives a set of patterns from a training corpus.**
- **Example of Rule generated using Crystal.**
 - **Conceptual Node for visiting-a-place-or-people event:**
 - **Verb: visited (active verb) (trigger word)**
 - **Visitor: V (person)**
 - **Has-location: P (place)**
 - **Start-time: ST (time-point)**
 - **End-time: ET (time-point)**
- **Example of patterns:**
 - **X visited Y on the date Z**
 - **X has been awarded Y money from Z**

- ◆ Badger makes instantiation of templates.
- ◆ In our example (David's Brown story), Badger instantiates the following slots of a Event -1 frame:
 - ◆ Type: visiting-a-pace-or-people
 - ◆ Place: The OU
 - ◆ Visitor: David Brown



```
(Def-instance visit-of-david-brown-the-chairman-of-the-university  
  visiting-a-place-or-people  
  ((start-time wed-15-oct-1997)  
   (end-time wed-15-oct-1997)  
   (has-location the-ou)  
   (visitor david-brown-the-chairman-of-the-university)  
  )  
)
```



- David Brown's story output after the OCML code is sent to Webonto.

Instance of **visiting-a-place-or-people**

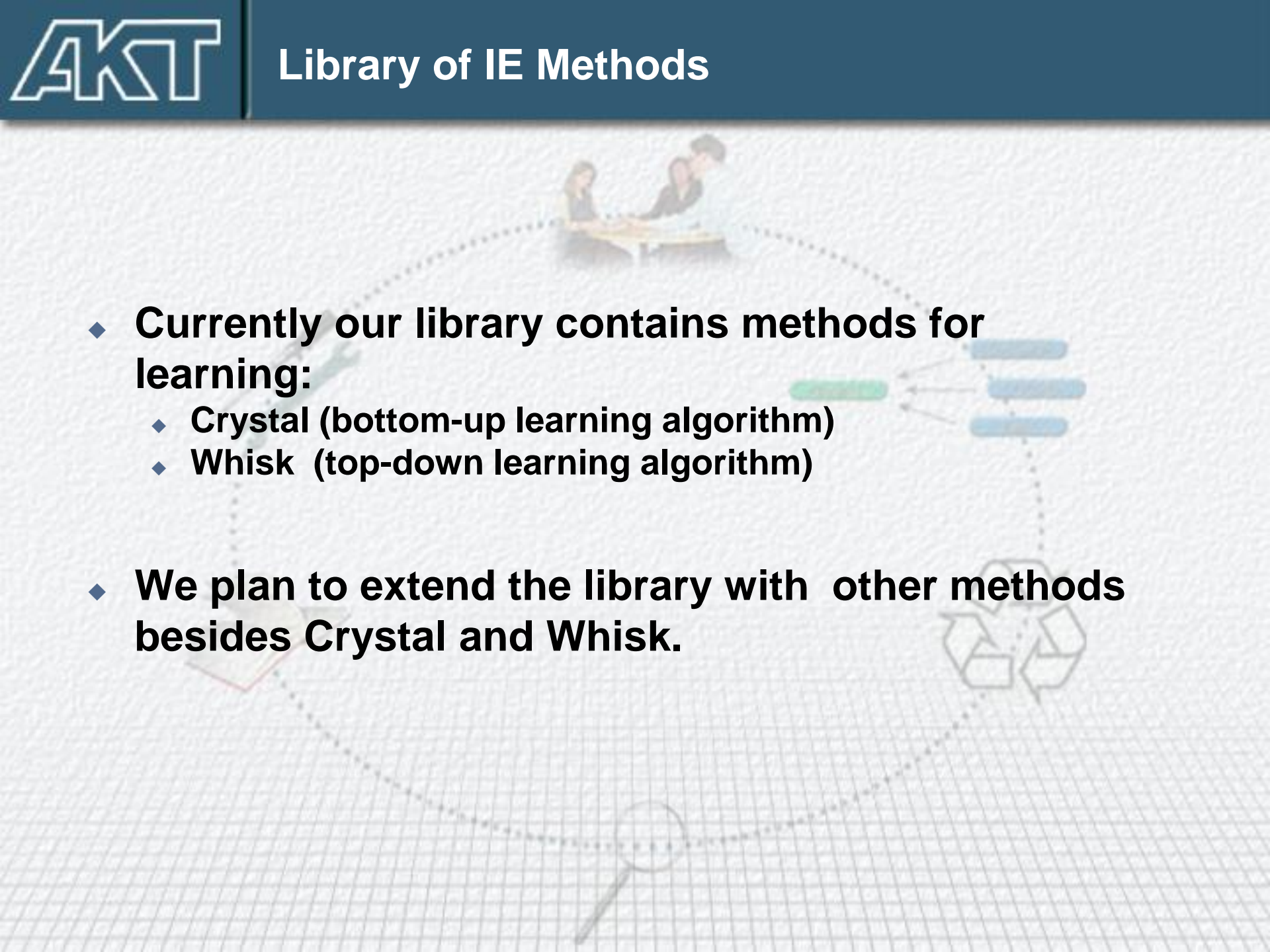
Name:

Click on a slot name to see examples of its use

has-duration	<input type="text" value="1-day"/>	duration	<input type="text" value="None"/>
start-time	<input type="text" value="wed-15-oct-1997"/>	time-point	<input type="text" value="None"/>
end-time	<input type="text" value="wed-15-oct-1997"/>	time-point	<input type="text" value="None"/>
has-location	<input type="text" value="the-ou"/>	location	<input type="text" value="None"/>
visitor	<input type="text" value="david-brown-the-chair"/>	(or person group-of-people)	<input type="text" value="None"/>
people-or-organization-being-visited	<input type="text" value=""/>	(or person organization)	<input type="text" value="None"/>


OK Cancel

Unsigned Java Applet Window

- ◆ **Currently our library contains methods for learning:**
 - ◆ **Crystal (bottom-up learning algorithm)**
 - ◆ **Whisk (top-down learning algorithm)**
 - ◆ **We plan to extend the library with other methods besides Crystal and Whisk.**
- 

- ◆ **Whisk: learns information extraction rules**
 - ◆ can be applied to semi-structured text (text is un-grammatical, telegraphic).
 - ◆ can be **applied** to free text (syntactically parsed text).
- ◆ **It uses a top-down induction algorithm seeded by a specific training example.**
- ◆ **Whisk has been used:**
 - ◆ **CNN weather forecast in HTML**
 - ◆ **BigBook addresses in HTML**
 - ◆ **Rental ads in HTML (our second domain)**
 - ◆ **Seminar announcements**
 - ◆ **job posting**
 - ◆ **Management succession text from MUC-6**

- ◆ **Domain Rental Adverts:**
- ◆ **Ballard - 2 Br/2 Ba, top flr, d/w 1000 sf, \$820. (206) 782-2843.**
- ◆ **Rule expressed as regular expression:**
- ◆ **ID 26 Pattern:: * (Nghbr) * (<digit>) 'Br' * '\$' (<number>).**
- ◆ **Output:: Rental{Neighbourhood \$1} {Bedrooms \$2} {Price \$3}**

- ◆ Items in green colour are semantic word classes.
 - ◆ Nghbr :: Ballard | Belltown| ...
 - ◆ digit :: 1|2|...|9
 - ◆ number :: (0-9)*
 - ◆ Complexity : restricted wild card therefore, time is not exponential.
- 
- The diagram shows a green horizontal bar on the left with three arrows pointing to three blue horizontal bars stacked vertically on the right. A dashed line extends from the right side of the blue bars, curving downwards and ending at a magnifying glass icon at the bottom center of the slide.

- ◆ We had built a tool which extracts knowledge using and Ontology, IE component and OCML pre-processor.
- ◆ We had worked with 2 different domains (KMi stories and Rental adverts)
 - ◆ first domain
 - ◆ Precision over 95%
 - ◆ second domain
 - ◆ Precision: 86% - 94%
 - ◆ Recall: 85% - 90%
- ◆ We will integrate more IE methods in our system.
- ◆ To extend our system in order to produce XML output, RDFS,...
- ◆ to integrate visualisation capabilities

