Semantic Web Tools

Based on the slides by Dipsy Kapoor
Outline

Semantic Web Tools

- Piggy Bank/Solvent
- Jena/ARQ
Piggy Bank

• “Semantic Web Browser” by SIMILE project at MIT

• Available as a Firefox plug-in at http://simile.mit.edu/piggy-bank/

• Can ingest RDF data directly from websites that contain meta information, and can execute screen-scrapers to extract content from normal websites and then convert that data to RDF

Warning: Piggy Bank works with Firefox 2.0.x only.
Piggy Bank

- Features of the browser:
  - Users can find, collect, annotate RDF
  - Search and faceted browse of local RDF
  - Export data in RDF/N3

But this tool is **REALLY USELESS** when websites have no RDF describing objects in them!
Solvent

http://simile.mit.edu/wiki/Solvent

- Screen-Scraper (javascript-based) developed by the SIMILE project
- An extension of Piggy Bank
- Extracts data and exports it as RDF/n3
Piggy Bank and Solvent Demo

Scrape data from http://boston.craigslist.org/aap/ and export to RDF
Objective: turn this data to RDF file
Activate Solvent
Select rows to scrape

Just crop the first row, then Solvent will do the rest (based on DOM structure)
Name each extracted attribute

Every item must have URI in URI format! (if not, Solvent will not scrape anything)
Custom properties

You can define custom properties.
Remember the name of a property should be in URI form
e.g. http://boston.craiglist.org/app#area
Generate the scraper

Test if the code can scrape the page

Javascript code for scrapping the page
Define type (row type) & URL pattern

```javascript
data.addStatement(uri, rdf + 'type', 'http://simile.mit.edu/ns#Unknown', false); // Use your own type here
```

Change to the type of the row (e.g. Apartment)

Define URL pattern – so Piggy Bank will use this scraper when the present URL matches to the scraper’s URL pattern.
Save your scraper & install it to Piggy Bank

Two files (scraper’s code & metadata) are generated

Then open the .n3 file (craiglist.n3) with Firefox. Piggy Bank will load this scraper on its environment (next slide).
Install the scraper to Piggy Bank

1. Click save to save this scraper to Piggy Bank repository

2. Activate the scraper
Scrape the Data!

1. Visit boston.craiglist.org/aap again then click at the coin icon

You have to restart your browser to have your new scraper works properly!!

Scraping process might take time! (don’t panic)
Export the data to RDF

Then save the RDF to somewhere on your machine
Jena

- A Semantic Web Framework that includes:
  - A RDF API
  - Reading and writing RDF in RDF/XML, N3 and N-Triples
  - An OWL API
  - In-memory and persistent storage
  - SPARQL query engine

- Documentation available at http://jena.sourceforge.net/documentation.html
ARQ

- ARQ is Jena’s implementation of SPARQL.
- ARQ documentation is available at: http://jena.sourceforge.net/ARQ/documentation.html
- Can write queries using the Jena Java API or can write them in text files and execute them using ARQ’s command line utility.
PREFIX SM: <http://simile.mit.edu/2005/05/ontologies/location#>
PREFIX DC: <http://purl.org/dc/elements/1.1/>

SELECT DISTINCT ?title ?address 
FROM <sources/starbucks.n3>
WHERE {
  ?x SM:address ?address .
}
PREFIX SM: <http://simile.mit.edu/2005/05/ontologies/location#>
PREFIX DC: <http://purl.org/dc/elements/1.1/>

SELECT DISTINCT ?title ?address
FROM <sources/starbucks.n3>
WHERE {
  ?x SM:address ?address .
  FILTER (regex(?address, "Figueroa", "i")) .
}

regex is a X-Path function. Other functions available at this tutorial -
http://jena.sourceforge.net/ARQ/Tutorial/
Querying multiple sources, Using simple JOIN

PREFIX SM: <http://simile.mit.edu/2005/05/ontologies/location#>
PREFIX DC: <http://purl.org/dc/elements/1.1/>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>

SELECT DISTINCT ?title ?zip ?temperature
FROM NAMED <sources/starbucks.n3>
FROM NAMED <sources/weather.n3>
WHERE {
  GRAPH <sources/starbucks.n3> {
  } .
  GRAPH <sources/weather.n3> {
  } .
}

More examples (might be useful for your hw):
Useful Links

- Solvent - [http://simile.mit.edu/solvent/](http://simile.mit.edu/solvent/)
- Jena Documentation - [http://jena.sourceforge.net/documentation.html](http://jena.sourceforge.net/documentation.html)
- ARQ Documentation - [http://jena.sourceforge.net/ARQ/documentation.html](http://jena.sourceforge.net/ARQ/documentation.html)
- Portable firefox (in case you don’t want to remove your firefox 3.0) - [http://portableapps.com/apps/internet/firefox_portable](http://portableapps.com/apps/internet/firefox_portable)