Semantic Technologies
Motivation and Standards
Overview

• Motivation
  – Web (of Documents)
  – Web of Data (Resources)

• Identifying resources - URIs
• Describing resources - RDF
• Retrieving descriptions - HTTP
• Querying descriptions - SPARQL
• Inference over:
  – (basic) descriptions – RDFS
  – (rich) descriptions - OWL
Motivation

• Consider yesterday’s news (March 15\textsuperscript{th}) :

• We’re used to accessing the page with a ‘URL’

• And used to navigating to other pages with a \textit{hyperlink} based on another \textit{URL}
Motivation

• Google crawl these documents to build a graph
Motivation

- This graph is useful to compute Page Rank

References

Motivation

- This graph is useful to compute Page Rank
- But this link:

  does *not* make it clear what relationship holds between the 2 documents:
  - “I agree with this”
  - “I refute this”
  - “I’ll show you where Hillsborough is”
Motivation

(Machine processable) semantics is about two things:

• Identifying all resources not just documents

  http://dbpedia.org/resource/Hillsborough_disaster
  http://dbpedia.org/resource/Hillsborough_Stadium
  http://dbpedia.org/resource/Sheffield

• Relating all resources not just documents explicitly and unambiguously

  http://www.bbc.co.uk/news/uk-17382896
  http://purl.org/dc/elements/1.1/subject
  http://dbpedia.org/page/Hillsborough_disaster
  http://dbpedia.org/page/Hillsborough_Stadium
  http://dbpedia.org/page/Sheffield

Semantic Technologies
Semantic Technology

To achieve this we need to:

- **Identify** – URIs = Uniform Resource Identifiers
  - http://dbpedia.org/resource/Sheffield

  Anything is allowed, but some recommendations
  - Use http:// URIs, because these can be looked up
  - Should be human-readable, natural hierarchy
    - e.g. http://mydomain.com/people/barry_bishop
  - Should be stable
    - should not change when technology changes
    - e.g. http://mydomain.com/people/barry.php

- ...
To achieve this we need to:

- **Identify** – URIs = Uniform Resource Identifiers
  
  http://dbpedia.org/resource/Sheffield

- **Describe** – RDF = Resource Description Framework

  http://dbpedia.org/resource/Hillsborough_Stadium
  http://dbpedia.org/ontology/location
  http://dbpedia.org/resource/Sheffield

  Each relationship is defined by one ‘triple’:

  ```
  <http://dbpedia.org/resource/Hillsborough_Stadium>
  <http://dbpedia.org/ontology/location>
  <http://dbpedia.org/resource/Sheffield> .
  ```

  ...
RDF Literals

- As well as relating URI-identified resources, we can attach ‘literal’ metadata to resources:

  `<http://dbpedia.org/resource/Sheffield>`
  `rdfs:label "Sheffield",`
  "셰필드"@ko;
  `dbpedia-owl:populationUrban "640720"^^xsd:integer;`
  `dbpedia-owl:populationUrbanDensity 3949.20.`

(Still just triples, abbreviated)
RDF Statements

- RDF Statements (triples) are formed of three parts:

1. Subject: This is the Resource that the statement is about: URI (or blank node)
2. Predicate: The property that relates the subject and object: URI
3. Object: The value of the property: URI (, blank node) or literal
RDF Graphs

- Resources can be both subjects and objects
- Literals can only be objects
- A collection of statements makes a directed graph
- RDF documents are easily combined
  - cf. Linked Open Data
RDF Integration

• Each resource may be the subject and object of statements from different sources

http://dbpedia.org/resource/Hillsborough_disaster

http://dbpedia.org/resource/Hillsborough_Stadium

http://dbpedia.org/resource/Sheffield

http://www.bbc.co.uk/news/uk-17382896

http://purl.org/dc/elements/1.1/subject

<http://www.bbc.co.uk/news/uk-17382896>
<http://purl.org/dc/elements/1.1/subject>

<http://dbpedia.org/resource/Hillsborough_disaster>
<http://dbpedia.org/ontology/data>

<http://dbpedia.org/resource/Hillsborough_Stadium>
<http://dbpedia.org/ontology/location>
<http://dbpedia.org/resource/Sheffield>.
Open Datasets

GeoNames

The GeoNames geographical database is available for download free of charge under a (Creative Commons) CC Attribution license. It contains over 10 million geographical names and consists of 7.5 million unique features.

<http://sws.geonames.org/7646085/>

gn:name "Hillsborough Stadium" ;
wgs84_pos:lat 53.41139 ;
wgs84_pos:long -1.50072 ;
gn:parentFeature

MusicBrainz

The MusicBrainz music metadata database is available for download under a public domain (CC no rights reserved) and CC Attribution-NonCommercial-ShareAlike 2.0 license. It contains over 10 million track descriptions.
To achieve this we need to:

- **Identify** – URIs = Uniform Resource Identifiers
  
  `http://dbpedia.org/resource/Sheffield`

- **Describe** – RDF = Resource Description Framework
  
  `<http://dbpedia.org/resource/Hillsborough_Stadium>`
  `<http://dbpedia.org/ontology/location>`
  `<http://dbpedia.org/resource/Sheffield>`

- **Retrieve** – HTTP = HyperText Transfer Protocol

- ...
HTTP

- HTML Retrieval

```
GET /training HTTP/1.1
Host: www.ontotext.com
Accept-Language: en
```

- Can negotiate on (human) language, but also...

```
HTTP/1.0 200 OK
Date: Thu, 15 Mar 2012 13:26:55 GMT
Content-Type: text/html

<html>
<head> ...
</head>
<body>
```
HTTP Content Negotiation

- HTML and RDF Retrieval

GET /resource/Sheffield HTTP/1.1
Host: dbpedia.org
Accept: text/html

HTTP/1.1 303 See Other
Date: Thu, 15 Mar 2012 13:37:48 GMT
Location: http://dbpedia.org/data/Sheffield.xml

GET /resource/Sheffield HTTP/1.1
Host: dbpedia.org
Accept: application/rdf+xml

HTTP/1.1 303 See Other
Date: Thu, 15 Mar 2012 13:35:14 GMT
Location: http://dbpedia.org/page/Sheffield
HTTP Retrieval of Hash URIs

- RDF Retrieval of description for
  http://www.bbc.co.uk/programmes/b00h0rnf#programme

GET /programmes/b00h0rnf HTTP/1.1
Host: www.bbc.co.uk
Accept:application/rdf+xml

HTTP/1.1 200 OK
Date: Thu, 15 Mar 2012 14:01:42 GMT

<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  ...>
  <po:Episode rdf:about="/programmes/b00h0rnf#programme">
    <po:pid>b00h0rnf</po:pid>
    <dc:title>01/01/09 (Double episode)</dc:title>
    <po:short_synopsis>Sean contacts Roxy and Stacey as the new year dawns. Whitney struggles to fit in.</po:short_synopsis>
  </po:Episode>
</rdf:RDF>
Semantic Technologies (cont.)

To achieve this we need to:

- **Identify** – URIs = Uniform Resource Identifiers
  
  http://dbpedia.org/resource/Sheffield

- **Describe** – RDF = Resource Description Framework

  ```
  <http://dbpedia.org/resource/Hillsborough_Stadium>
  <http://dbpedia.org/ontology/location>
  <http://dbpedia.org/resource/Sheffield> .
  ```

- **Retrieve** – HTTP = HyperText Transfer Protocol

- **Query** – SPARQL = Protocol and RDF Query Language

  dbpedia:Hillsborough_disaster dbont:location ?location

- ...
What is SPARQL?

- SPARQL is a query language for RDF data
  - http://www.w3.org/TR/rdf-sparql-query/
- It is also a protocol based on HTTP
- It is designed around graph patterns
  - Graph patterns use (Turtle) syntax (seen earlier)
- SPARQL 1.0 is query only
- SPARQL 1.1 includes syntax for updates
SPARQL Example

```
SELECT ?title (COUNT(?ep) AS ?episodes)
WHERE {?brand po:genre <http://www.bbc.co.uk/programmes/genres/drama/soaps#genre>;
       dc:title ?title ;
       po:episode ?ep}
GROUP BY ?brand
ORDER BY DESC(?episodes)
LIMIT 10
```

<table>
<thead>
<tr>
<th>title</th>
<th>episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EastEnders</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Semantic Technologies (cont.)

To achieve this we need to:

- **Identify** – URIs = Uniform Resource Identifiers
  
  http://dbpedia.org/resource/Sheffield

- **Describe** – RDF = Resource Description Framework
  
  <http://dbpedia.org/resource/Hillsborough_Stadium>
  <http://dbpedia.org/ontology/location>
  <http://dbpedia.org/resource/Sheffield> .

- **Retrieve** – HTTP = HyperText Transfer Protocol

- **Query** – SPARQL = Protocol and RDF Query Language
  
  dbpedia:Hillsborough_disaster dbont:location ?location

- **Infer**
  
  – RDFS = RDF Schema Language
  – OWL = Web Ontology Language
RDF Schema allows us to define

- **Classes**
  - group resources:
    - specialise/generalise:
      - type
        - http://purl.org/ontology/po/Episode
        - http://purl.org/ontology/po/Programme
      - subClassOf
        - http://purl.org/ontology/po/Programme

- **Properties**
  - define relationships:
    - po:episode rdf:type rdfs:Property ;
      - rdfs:domain po:Brand ;
      - rdfs:range po:Episode .
RDF Schema

Programmes ontology

- Defines
  - class hierarchy
  - properties

using RDFS, allowing RDF via HTTP, and SPARQL queries (as seen)
• Allows more complex modelling
  – Constraints
    • Resources in a given class *must* have certain properties
    • E.g. Cardinality on properties
  – Intentional classes
    • Under certain properties resources are inferred into a class
    • E.g. Every programme with a broadcast date is an episode
  – Intentional properties
    • Under certain conditions a property is inferred to hold between two resources
    • E.g. If X is Y’s parent’s male sibling, X is Y’s uncle
Conclusion

• The Web (of documents) is generalised by
  – Applying URIs to non-document resources
  – Using RDF to describe relationships
  – Providing these descriptions using HTTP
  – Allowing querying over these descriptions using SPARQL
  – Inferring implicit relationships using RDFS and OWL

we call the result the public exposure of these the Web of Data

• Linked Data principles can equally be applied to form applications within the enterprise