AKTivePSI

An Example of Semantic Web Data Integration for Government

Professor Nigel Shadbolt
University of Southampton
Using advanced knowledge management technology to improve the delivery of policy and public services across Government
Making the Web
Semantic...

WWW 2002
THE ELEVENTH INTERNATIONAL
WORLD WIDE WEB CONFERENCE
Sheraton Waikiki Hotel
Honolulu, Hawaii, USA
7-11 May 2002

1 LOCATION. 5 DAYS. LEARN. INTERACT.

Registered participants coming from:

Australia • Canada • Chile • Denmark • France • Germany • Ghana • Hong Kong • India • Italy • Ireland • Japan • Malta • New Zealand • The Netherlands • Norway • Singapore • Switzerland • The United States • Vietnam • Zambia

On 7-11 May 2002, Honolulu, Hawaii will provide the backdrop for The Eleventh International World Wide Web Conference. This prestigious series of the International World Wide Web Conference Committee (W3C²) attracts participants from around the world, and it provides a public forum for the World Wide Web Consortium (W3C) through the annual W3C track.

The conference is being organized by the International World Wide Web Conference Committee (W3C²), the University of Hawaii and the Pacific Telecommunications Council (PTC).

FEATURED SPEAKERS (CONFIRMED)

Tim Berners-Lee, inventor of the World Wide Web and Director of the W3C, who now holds the 3Com Founders chair at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT).

Richard A. DeMillo, vice president and chief technology officer for Hewlett-Packard Company.

Ian Foster, guru of “Grid Computing”, associate professor at the University of Chicago. McArthur Prize Winner.
via ontologies...

This is a type of object event and this is its title.

This is the URL of the web page for the event.

This is a type of object photograph and the photograph is of Tim Berners-Lee.

Tim Berners-Lee is an invited speaker at the event.

<owl:Class rdf:ID="Conference">
  <rdfs:subClassOf rdf:resource="#Meeting-Taking-Place"/>
  <rdfs:subClassOf rdf:resource="#Publication-Type-Event"/>
  -<rdfs:subClassOf>
    -<owl:Restriction>
      <owl:onProperty rdf:resource="#published-proceedings"/>
      <owl:allValuesFrom rdf:resource="#Conference-Proceedings-Reference"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
can semantically enrich anything…

- Web documents
- Databases
- Scientific structures
- Workflow
- Publications
- People
• Content harvested from multiple heterogeneous sources
  
  – Higher Education directories
  – 2001 RAE submissions
  – UK EPSRC project database
  – Info on personnel, projects and publications harvested for 5 or 5* CS departments in the UK
Other examples

- e-health
- e-defence
Aims of AKTivePSI

• Show how information in existing databases can be made available in scalable **semantic knowledge bases**
  – Using semantic web languages to represent and query the data

• Show how all this data can be linked to create an extended knowledge network

• Show how **ontologies** can represent the given data

• Demonstrate examples of **added value**

• Investigate the suitability of **IPSV** for representing government data

• Identify **knowledge gaps** between existing databases, and how such gaps can be filled

• Other Opportunities
Key Working Assumptions

• Decided to follow an approach that simulates a real-life scenario
  – Minimum disruption to existing data flows and models
  – With minimum or no cost to the participants

• One dataset at a time
  – No preparations are needed
  – Give us the data as it is, in any format and delivery method

• Convert databases into focused ontologies using simple scripts
  – Use as much automation as possible to extract the necessary metadata from existing databases and documents
  – No data is to be handled manually!
    • E.g. when inserting into knowledge bases, linking to other data, merging duplications

• Practical ontologies!
  – Keep the ontologies small and manageable whenever possible
  – Ontologies are to be constructed to represent the data in a given database, not to represent an entire domain
  – Larger ontologies will be required later for integration
    • Not something to worry about from the start
Trust and Provenance

• It is important to maintain the provenance of the data we collect

• Each dataset is stored in a separate Knowledge Base, using a dedicated ontology
  – E.g., Camden would have its own knowledge base, and Lewisham would have theirs
  – To minimise risk of contaminating one dataset with another
  – To make sure that the source of the data can be fully traced

• Each ontology clearly shows who provided the data and when
  – We can also represent who is the data owner, distributor, creator, etc
  – The data in the KB sometimes directly points to its source

• Ontologies are separate, but mapped/linked to each other
Creating the Knowledge Bases

1. Study the data
2. Build the ontologies
3. Convert data into RDF
4. Write conversion scripts
5. Assert into KB
Datasets

• Camden Council
  – Land & Property Gztt.
  – Food premises
  – Local Businesses
  – Licences
  – Councillors and Committees
  – Some meeting minutes

• Lewisham Council
  – Land & Property Gztt.
  – LBL

• Ordnance Survey
  – Points of Interest
  – MasterMap
  – Address Layers 1 and 2

• London Gazette
  – All database records 1998 onwards
Camden Borough Council
Camden’s LPG Dataset

- Camden has provided the Land & Property Gazetteer
  - Contains info about properties in Camden, full address, coordinates, flag for residential/non-residential/mixed.
  - Provided as a CSV file
  - Contains over 125K records

<table>
<thead>
<tr>
<th>UID</th>
<th>Desc</th>
<th>UPRN</th>
<th>ParentUPRN</th>
<th>UNIT</th>
<th>BuildingName</th>
<th>Street</th>
<th>Postcode</th>
<th>Town</th>
<th>EASTING</th>
<th>NORTHING</th>
</tr>
</thead>
<tbody>
<tr>
<td>45526</td>
<td>Residential (Unit)</td>
<td>5014096</td>
<td>5105271</td>
<td>Flat 43</td>
<td>Holly Lodge Mansions</td>
<td>Oakeshott Avenue</td>
<td>N6 6DS</td>
<td>London</td>
<td>528418.8</td>
<td>186856.3</td>
</tr>
<tr>
<td>45532</td>
<td>Residential (Unit)</td>
<td>5014102</td>
<td>5105271</td>
<td>Flat 29</td>
<td>Holly Lodge Mansions</td>
<td>Oakeshott Avenue</td>
<td>N6 6DS</td>
<td>London</td>
<td>528418.8</td>
<td>186856.3</td>
</tr>
<tr>
<td>45533</td>
<td>Residential (Unit)</td>
<td>5014103</td>
<td>5105271</td>
<td>Flat 28</td>
<td>Holly Lodge Mansions</td>
<td>Oakeshott Avenue</td>
<td>N6 6DS</td>
<td>London</td>
<td>528418.8</td>
<td>186856.3</td>
</tr>
<tr>
<td>45538</td>
<td>Residential (Unit)</td>
<td>5014108</td>
<td>5105271</td>
<td>Flat 51</td>
<td>Holly Lodge Mansions</td>
<td>Oakeshott Avenue</td>
<td>N6 6DS</td>
<td>London</td>
<td>528418.8</td>
<td>186856.3</td>
</tr>
</tbody>
</table>

- Does not include more info for non-residential or mixed properties, eg type of business
Camden LPG Ontology

26 Concepts

9 Object properties
- Links between concepts
- eg Address --has-post-code--> PostCode

17 Datatype properties
- Links between concepts and non-concepts (eg strings, numbers)
- Eg PostCode --post-code--> String

Ontology built using namespace
http://www.camden.gov.uk/propertyOntology
- Easy to trace URI in knowledge base to it’s origin

Produced 2.3 million RDF triples
LPG Data Conversion

@prefix : <http://www.camden.gov.uk/propertyOntology#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.

:infoSourceCamdenLondon :has-source "Camden’s Land and Property Gazetteer".

:property5014096 rdf:type :Flat.
:property5014096 :local-ID "46526".
:property5014096 :has-UPRN "5014096".
:property5014096 :has-address
  [ :has-building-name "Holly Lodge Mansions";
    :has-building-number "";
    :has-unit "Flat 43";
    :has-post-code :postcodeN66DS].

:street20400028 rdf:type :Street.
:street20400028 :street-name "Oakeshott Avenue".
:street20400028 :has-USRN "20400028".

:postcodeN66DS :post-code "N6 6DS".

:east528418.8north186856.3 rdf:has-type :EastingNorthing.
:east528418.8north186856.3 :has-easting "528418.8".
:east528418.8north186856.3 :has-northing "186856.3".
:property5014096 :has-coordinates :east528418.8north186856.3.
Food Premises

- Dataset contains information on premises in Camden that produces, handles, or serves food
  - E.g. restaurants, schools kitchens, canteens

- Includes business name, results of last food hygiene and standards inspection checks, addresses, premises type (e.g. restaurant, school, bar)

- Data provided in xls spreadsheet,
  - 2.8 thousand records
  - Produced over 84K RDF triples
  - Ontology stats: 165 classes, 17 object properties, 15 datatype properties
Ordnance Survey
• Data provided:
  – Master maps for Camden and Lewisham
    • GIS maps showing land boundaries and borders
  – Address Layer 1
    • Data about buildings, addresses, coordinates
  – Address Layer 2
    • Buildings are classified into types (e.g., hospital, university, hotel)
  – Ontology for Address Layer dataset (osgb.owl)
    • Written in OWL to represent the data in this dataset
• Added minor extensions to osgb.owl
  – To represent few extra concepts
  – To facilitate mapping to other ontologies

• Converted this xml dataset to RDF and stored in 3Store against the extended OS ontology
  – Produced 758 thousand triples
  – Mainly buildings, addresses, and coordinates
OS Ontology

- 93 Concepts
- 80 Object properties
  - Links between concepts
- 72 Datatype properties
  - Links between concepts and non-concepts (e.g., strings, numbers)
- Ontology built using namespace [http://www.ordnancesurvey.co.uk/xml/namespaces/osgb](http://www.ordnancesurvey.co.uk/xml/namespaces/osgb)
  - Extends a standard ontology [http://www.opengis.net/gml](http://www.opengis.net/gml)
<osgb:addressPointMember>
  <osgb:AddressPoint rdf:ID="osgb1000002148475495">
    <osgb:version>1</osgb:version>
    <osgb:versionDate>2002-08-17</osgb:versionDate>
    <osgb:theme>Address</osgb:theme>
    <osgb:addressStatus>
      <osgb:matchStatus>Matched</osgb:matchStatus>
      <osgb:physicalStatus>Existing</osgb:physicalStatus>
      <osgb:positionalQuality accuarcy="Surveyed">Final</osgb:positionalQuality>
      <osgb:structureType>Permanent Building</osgb:structureType>
    </osgb:addressStatus>
    <osgb:OSAPR>APKEBA8H52C45WT0QT</osgb:OSAPR>
    <osgb:point>
      <gml:Point srsName="osgb:BNG">
        <gml:coordinates>537960.700,176636.300</gml:coordinates>
      </gml:Point>
    </osgb:point>
    <osgb:postalAddress>
      <osgb:buildingNumber>1</osgb:buildingNumber>
      <osgb:thoroughfare>ROBINSKROFT MEWS</osgb:thoroughfare>
      <osgb:postTown>LONDON</osgb:postTown>
      <osgb:deliveryPointSuffix>1A</osgb:deliveryPointSuffix>
    </osgb:postalAddress>
    <osgb:postalAddressDate>2002-07-19</osgb:postalAddressDate>
    <osgb:reference1>1</osgb:reference1>
  </osgb:AddressPoint>
</osgb:addressPointMember>

conversion scripts

:AddressPointosgb1000002148475495
  :hasPostalAddress
  :PostalAddressosgb1000002148475495
  :PostalAddressosgb1000002148475495 rdf:type
  :PostalAddress.
  :PostalAddressosgb1000002148475495
    :hasDeliveryPointSuffix '1A'.
  :PostalAddressosgb1000002148475495
    :hasThoroughfare 'ROBINSKROFT MEWS'.
  :PostalAddressosgb1000002148475495 :hasRoad
    [rdf:type :Thoroughfare; :hasThoroughfareName
    'ROBINSKROFT MEWS'].
  :PostalAddressosgb1000002148475495
    :hasPostTown :TownLONDON.
  :TownLONDON rdf:type :Town.
  :TownLONDON :hasName 'LONDON'.
  :PostalAddressosgb1000002148475495
    :hasPostCode :PostCodeSE10_8DN.
  :PostCodeSE10_8DN :hasPostCodeValue 'SE10 8DN'.
  :PostalAddressosgb1000002148475495
    :hasCoordinates
    :PostalCoordinatesosgb1000002148475495.
  :PostalCoordinatesosgb1000002148475495 rdf:type
  gml:Coordinates.
  :PostalCoordinatesosgb1000002148475495
    gml:hasDecimal '537960.700,176636.300'.
Address Layer 2

- Similar to Address Layer 1, but with more place related information
  - E.g. name and category (hospital, school)

- Provided in xml format
  - Contained information about around 35K places
  - Converted into 11.7M RDF triples in 3Store
  - Ontology has 98 classes, 84 object properties, and 89 datatype properties
  - Most of this ontology is not used or needed for this data, but are inherited from standard geographical representations
Points of Interest Data
PointX, founded in 2001, is a joint venture company owned by the Ordnance Survey and Landmark Information Group.

PointX offers a “Comprehensive, up-to-date and accurate Points of Interest data for Great Britain“

Distributed by OS

Relies on various data supplier
- Eg OS, thomsonlocal.com, experian.com, and many others

OS provided PointX data for Camden and Lewisham
- Over 22.5 thousand records
- Create an ontology for PointX with 10 classes and 24 properties
- Produced nearly 467 thousand RDF triples
Points of Interest data

- Cafés, restaurants, hotels, bars, etc
- Full addresses
- Classification number indicating type of business
London Gazette
The entire LG is made available for this project
- Contains all the info since they started digitising their data in 1998
- Large database, many different type of notices
- Current data structure is difficult to parse
- TSO is currently redesigning the database

We focussed on insolvency and deceased person notices
- So far, we converted 4550 Appointment of Administrator notices for Corporate Insolvencies
  - However, many of the addresses were not parsed correctly!
  - Actual address of businesses are usually not available
    - Historical council data might be useful to fill this gap
  - Resulted in 120 thousand RDF statement
- For the deceased person data
  - 3.2 million RDF statements were created
Insolvency Ontology

- Ontology to represent notices of type Appointment of Administrator
- 22 Concept, 24 properties

• Integrating or mapping ontologies together improves cross-KB querying and understanding
  – But is not necessary to utilise the data

• Each ontology can be linked:
  – Directly to other local ontologies
  – Directly to other external ontologies, or via a shared reference ontology (such as IPSV)

• No need to be restricted to any given standard taxonomies
  – Such standards can never detail all types of data!
  – Use your own ontologies to represent your own data
    • Or reuse or modify an existing one to fit your data
  – Map your ontologies to the standards (eg IPSV)
Integration

• Three types of mapping and integration was applied:
  – Mapping of ontologies
    • Using CROSI – an AKT ontology mapping tool
  – Mapping of instance data
    • Scripts to search for duplications
    • Insertion of owl:sameAs in 3Store to link duplicated objects
  – Mapping of ontologies to IPSV
    • Had to be done manually
• Integrated Public Sector Vocabulary

• “IPSV now covers internal-facing as well as public-oriented topics”

• “Stay with IPSV if your purpose is to populate Subject metadata”

• 3080 preferred terms and 4843 non-preferred
Observations on IPSV

• Mainly designed to represent “topics” not “data”
  – Good for describing documents
  – Bad for describing data!
    • You can find metadata about housing topics, but there isn’t a ‘House’ class

• Not enough comments are given to explain the choice and meaning of Terms

• Some topics are scattered in many places
  – E.g licences are placed in many different IPSV branches

• The taxonomy can not be used as class hierarchy
  – Causes problems when using RDF/OWL inference
  – IPSV isA relations are for topics, not concepts

• Mapping to IPSV or similar broad ontologies is useful
  – Facilitates integration of distributed KBs
  – Helps to disambiguate local terminology
    • Eg insolvencyOntology:Court → ipsv:Courts of law (ie not a tennis court!)
    • foodPremises:Alternative_Medicine → ipsv:Complementary medecine

• Not enough abstract terms in IPSV
  – eg no term to represent Road or Street, but it has 15 road related terms, such as Road Accident, Road Works, Road Signs, Road Safety
  – nothing to map Addresses to
Example Mashups

1. Camden’s food premises + OS Address Layer 2 + PointX

2. Lewisham’s Land & Property Gazetteer + Address Layer 2 + PointX
Camden Food Premises

- Food premises db provides food hygiene check results
  - But does not have coordinates
- This was *mashed up* with AddressLayer 2 and PointX to retrieve coordinates
- Result is a map with locations of food premises in Camden, coloured according to their total score of hygiene
Public Awareness

- Received very good reviews from the public
  - Scoring 9.3 Food, 8.2 Service, 8.2 Atmosphere, and 9 Value out of 10 in [http://www.london-eating.co.uk/](http://www.london-eating.co.uk/)

- While it scored quite badly in Camden’s health checks:

| has-name                                      | has-potential-hazard-handling-value | has-potential-hazard-method-value | has-consumers-at-risk-value | has-hygiene-and-safety-compliance-value | has-structural-compliance-value | has-confidence-in-management-value | hasE-coli-0157Risk | has-risk-band | has-risk-total | has-telephone-number | date-last-review |
|-----------------------------------------------|-------------------------------------|------------------------------------|-----------------------------|----------------------------------------|---------------------------------|-------------------------------|-------------------|---------------|----------------|----------------------|-----------------|-----------------|
| Preparation High                              | LowRiskActivity                     | Few                                | ImprovementNeeded           | ImprovementNeeded                     | ImprovementNeeded               | ImprovementNeeded            | NotSignificant    | A             | 100           |                      | 14/04/2008       |

- Easy access to these results can act as a great incentive for businesses to stay “clean”
Lewisham’s LPG

- LPG gives addresses and coordinates of Lewisham properties
  - But has no information about what the property is (e.g. residential, business, restaurant)
- Mashed up with AddressLayer 2 and PointX to retrieved more info about the property
Some Observations

• Lack of temporal data
  – E.g. when a business was established, closed down

• No detail for why a record was changed
  – E.g. some dbs have dates of changes, but not clear what has changed

• No commonly used unique property numbers:
  – E.g. Bento Café:
    • PointX ID: 21012114
    • Camden UPRN: 5087738
    • OS2 UDPRN: 17647957
    • OS address key: 27172769

• Data does not distinguish between single and multi business premises
  – Camden food premises:
    • Bento Café 9 NW1 7PG
  – OS2
    • Bento Café 9 NW1 7PG
  – PointX
    • Bento Café 9 NW1 7PG, Perennis Ltd 9 NW1 7PG
  – Is this an error? Is it a business that replaced/ got replaced by Bento Café? Is it a company that is located above Bento Cafe and using the same address?
  – Answer: it is a large building with several businesses!
  – Perennis Ltd is not in any of our Camden’s datasets
Conclusions

• Small ontologies can do the job
  – Ontologies to limited domains
  – Can be integrated in various ways

• Use of ontologies
  – Data mapping and integration made easier
  – Helped to understand the data models
  – Flexibility of representation
  – Overall, we created around 19 million RDF statements

• Much can be gained when the data is integrated
  – Data about the same place or object is distributed across several databases and organisations
  – Data enrichment, consistency checks, better analysis
  – Better to integrate data from various sources, rather to duplicate it!

• Data access can be made easier
  – Mashups can be generated relatively easily
  – Search and retrieval across databases
  – Data can be published in “machine understandable” formats
We now have the key to

“unlocking the potential of public sector information”
Power and Insight about your Digital Identity

www.garlik.com
The Management Team

Tom Ilube, Chief Executive Officer
Previously, Tom was Chief Information Officer & Executive Committee member of Egg plc, the world’s largest pure online bank.

Mike Harris, Executive Chairman
As founding Chief Executive Officer of Egg plc, Mike took Egg from concept to a £1bn public company within 3 years.

Nigel Shadbolt, Chief Technology Officer
Nigel is a Professor in the School of Electronics & Computer Science at University of Southampton. His current research focus is the Semantic Web.
The Opportunity

- The internet is part of everyday life
  - 64% UK adults online
  - 19 MM have broadband access

- Web risks becoming a victim of its own success
  - 25 billion web pages & 100 billion by 2010
  - Will we be able to find what or who we are looking for?

- Private information exposed for others to use or abuse
  - Information you share or others share about you
  - Company or government data

- Big – and growing – issue for consumers
  - 39% concerned about their privacy on the internet
  - 35% don’t trust companies with the personal data they collect
Data Patrol - Overview

- Sensitive data:
  - Name: Prof Nigel Shadbolt
  - Address: 999
  - Conference Place:
  - Date of Birth:
  - Mother's Maiden Name:

- Credit profile:
  - Credit Rating:
  - Credit Score:
  - Income Rating: unknown
  - LifeStage Rating: unknown
  - Average House Price in Your Area: N/A

- Connections:
  - You have relationships with 0 people.
  - You are connected to 0 companies or organisations.
<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Further Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan Rouge</td>
<td>cleeseor.ist.psu.edu</td>
<td></td>
</tr>
<tr>
<td>Alex Vaughan</td>
<td>Companies House</td>
<td>Ecs Partners Limited</td>
</tr>
<tr>
<td>Amanda Hill</td>
<td>Companies House</td>
<td>Ecs Partners Limited</td>
</tr>
<tr>
<td>Andrew Brown</td>
<td>Companies House</td>
<td>Ecs Partners Limited</td>
</tr>
<tr>
<td>Arnold Perrington</td>
<td>Companies House</td>
<td>Ecs Partners Limited</td>
</tr>
<tr>
<td>Arthur Statt</td>
<td>cleeseor.ist.psu.edu</td>
<td></td>
</tr>
<tr>
<td>Beth Crandall</td>
<td>cleeseor.ist.psu.edu</td>
<td></td>
</tr>
<tr>
<td>Bo Hu</td>
<td>cleeseor.ist.psu.edu</td>
<td></td>
</tr>
<tr>
<td>Bob Weinga</td>
<td>cleeseor.ist.psu.edu</td>
<td></td>
</tr>
<tr>
<td>Catherine Goldsmith</td>
<td>Birtha, Marriages and Divorces</td>
<td></td>
</tr>
<tr>
<td>Christopher Brewster</td>
<td>informatik.uni-erlangen.de</td>
<td></td>
</tr>
<tr>
<td>Claire Entwistle</td>
<td>Companies House</td>
<td>Ecs Partners Limited</td>
</tr>
<tr>
<td>David De</td>
<td>cleeseor.ist.psu.edu</td>
<td></td>
</tr>
<tr>
<td>Ecs Partners Limited</td>
<td>Companies House</td>
<td></td>
</tr>
<tr>
<td>Semnet Limited</td>
<td>Companies House</td>
<td></td>
</tr>
<tr>
<td>Vista</td>
<td>www-users.cs.york.ac.uk</td>
<td></td>
</tr>
</tbody>
</table>