Data.dcs: Converting Legacy Data into Linked Data

Matthew Rowe

Organisations, Information and Knowledge Group

University of Sheffield

Outline

Problem

Legacy data contained within the Department of Computer Science

Motivation

Why produce linked data?

Converting Legacy Data into Linked Data:

Triplification of Legacy Data

Coreference Resolution

Linking into the Web of Linked Data

Deployment

Conclusions

Problem

- The Department of Computer Science (http://www.dcs.shef.ac.uk) provides a web site containing important legacy data describing

People

Research groups

Publications

Legacy data is defined as important information which is stored in proprietary formats

Each member of the DCS maintains his/her own web page

Heterogeneous formatting

Different presentation of content

Devoid of any semantic markup

Motivation

Leveraging legacy data from the DCS in a machine-readable and consistent form would allow related information to be linked together

People would be linked with their publications

Research groups would be linked to their members

Co-authors of papers could be found

Linking DCS data into the Web of Linked Data would allow additional information to be provided:

Listing conferences which DCS members have attended

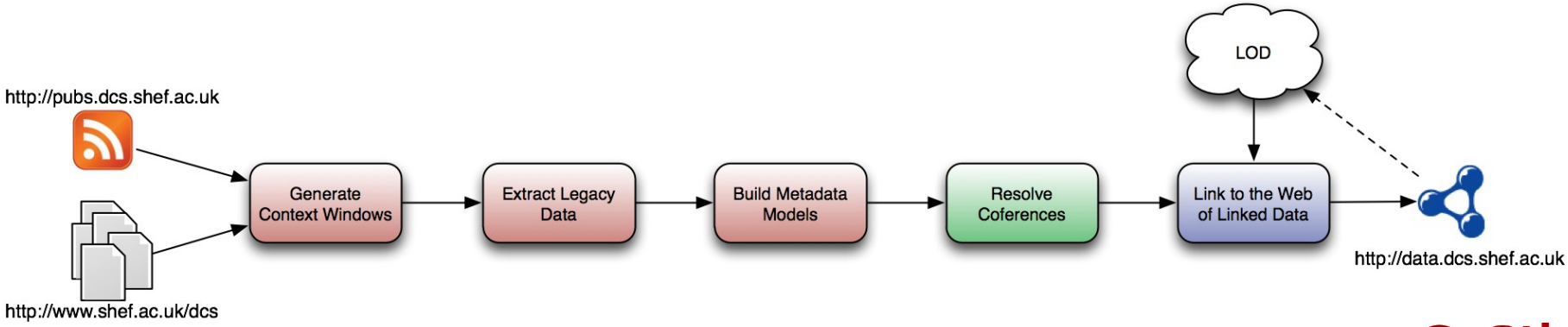
Provide up-to-date publication listings

- Via external linked datasets

Converting Legacy Data into Linked Data

The approach is divided into 3 different stages:

1. Triplification
    2. Converting legacy data into RDF triples
3. Coreference Resolution
    4. Identifying coreferring entities into the RDF dataset
5. Linked to the Web of Linked Data



Triplification of Legacy Data

The DCS publication database provides publication listings as XML

However, all publication information is contained within the same <description> element (title, author, year, book title):

```
<description>
  <![CDATA[Rowe, M. (2009). Interlinking Distributed Social Graphs.
  In <i>Proceedings of Linked Data on the Web Workshop, WWW 2009,
  Madrid, Spain. (2009)</i>. Madrid, Madrid, Spain.<br>
  <br>Edited by Sarah Duffy on Tue, 08 Dec 2009 09:31:30 +0000.]]>
</description>
```

DCS web site provides person information and research group listings in HTML documents

Information is devoid of markup and is provided in a heterogeneous formats

Triplification of Legacy Data

Vitaveska Lanfranchi

Research Associate

http://www.dcs.shef.ac.uk/~vita/
V.Lanfranchi@dcs.shef.ac.uk

Researcher on Human Computer Interaction, coordinator of the application area

Suvodeep mazumdar

Research Assistant

http://www.dcs.shef.ac.uk/~suvodeep/
S.Mazumdar@dcs.shef.ac.uk

Researching Human Computer Interaction and Visulisation

Matthew Rowe

Ph.D. Student

http://www.dcs.shef.ac.uk/~mrowe/
M.Rowe@dcs.shef.ac.uk

Researching Identity Disambiguation and Web 2.0

Lei Xia

Research Associate

http://www.dcs.shef.ac.uk/~lei/
l.xia@dcs.shef.ac.uk

Researching Information Extraction from Text

Click to edit Master text styles
Second level
- Third level
- Fourth level
- Fifth level

```
<div id="vita">
    <h4>Vitaveska Lanfranchi</h4>
    <img src="../images/people/vita.jpg"/>
    <p class="position">Research Associate</p>
    <ul>
        <li><a href="http://www.dcs.shef.ac.uk/~vita/">http://www.dcs.shef.ac.uk/~vita/</a></li>
        <li><a href="mailto:V.Lanfranchi@dcs.shef.ac.uk">V.Lanfranchi@dcs.shef.ac.uk</a></li>
    </ul>
    <p>Researcher on Human Computer Interaction, coordinator of the application area</p>
</div>
<div>
    <h4>Suvodeep mazumdar</h4>
    <img src="../images/deep.jpg"/>
    <p class="position">Research Assistant</p>
    <ul>
        <li><a href="http://www.dcs.shef.ac.uk/~suvodeep/">http://www.dcs.shef.ac.uk/~suvodeep/</a></li>
        <li><a href="mailto:S.Mazumdar@dcs.shef.ac.uk" property="foaf:mbox">S.Mazumdar@dcs.shef.ac.uk</a></li>
    </ul>
    <p>Researching Human Computer Interaction and Visulisation</p>
</div>
<div>
    <h4>Matthew Rowe</h4>
    <img src="../images/people/matt.jpg"/>
    <p class="position">Ph.D. Student</p>
    <ul>
        <li><a href="http://www.dcs.shef.ac.uk/~mrowe/">http://www.dcs.shef.ac.uk/~mrowe/</a></li>
        <li><a href="mailto:m.rowe@dcs.shef.ac.uk">M.Rowe@dcs.shef.ac.uk</a></li>
    </ul>
    <p>Researching Identity Disambiguation and Web 2.0</p>
</div>
<div>
    <h4>Lei Xia</h4>
    <img src="../images/people/lei.png"/>
    <p class="position">Research Associate</p>
    <ul>
        <li><a href="http://www.dcs.shef.ac.uk/~lei/">http://www.dcs.shef.ac.uk/~lei/</a></li>
        <li><a href="mailto:l.xia@dcs.shef.ac.uk">l.xia@dcs.shef.ac.uk</a></li>
    </ul>
    <p>Researching Information Extraction from Text</p>
</div>
```

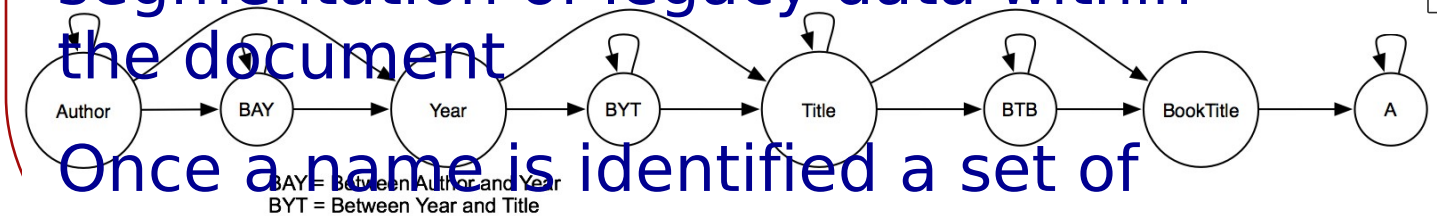
Triplification of Legacy Data

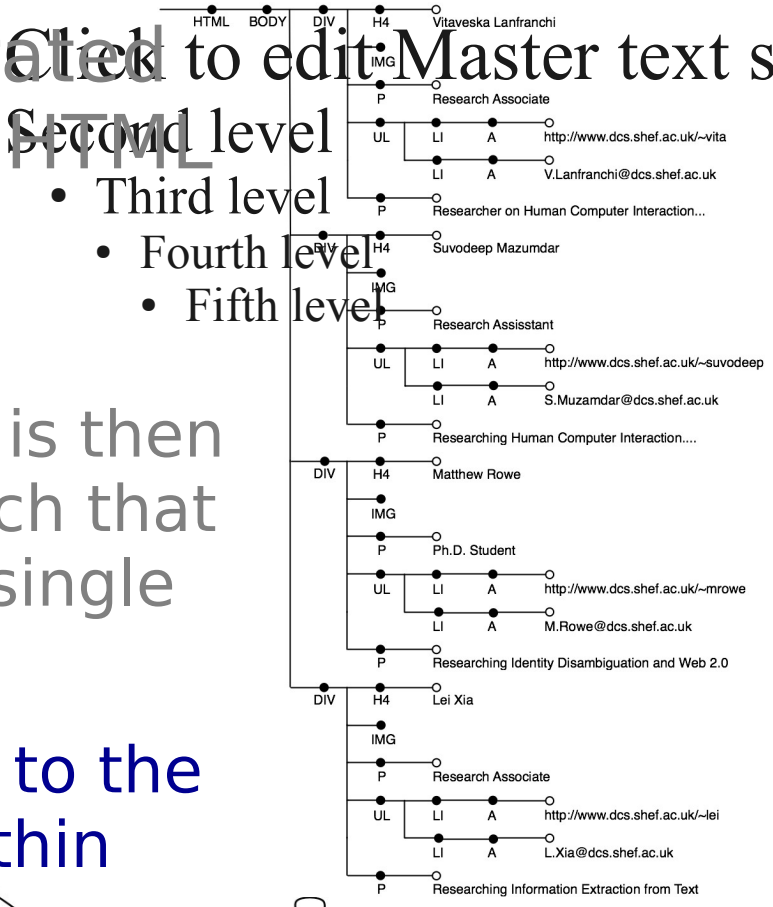
Context windows are generated by identifying portions of a HTML document which contain a person's name

The structure of the HTML DOM is then used to partition the window such that it contains information about a single person

HTML markup provides clues as to the segmentation of legacy data within the document

Once a name is identified a set of algorithms moves up the DOM tree until a layout element is discovered





Triplification of Legacy Data

An RDF dataset is built from the extracted legacy data

This provides the source dataset from which a linked dataset is built

For person information triples are formed as follows:

```
<http://data.dcs.shef.ac.uk/person/12025>
  rdf:type foaf:Person ;
  foaf:name "Matthew Rowe" .
<http://www.dcs.shef.ac.uk/~mrowe/publications.html>
  foaf:topic
<http://data.dcs.shef.ac.uk/person/12025>
```

Coreference Resolution

The triplification of legacy data contained within the DCS web sites (from ~12,000 HTML documents) produced 17,896 instances of foaf:Person and 1,088 instances of bib:Entry

Contains many equivalent foaf:Person instances

Must also assign people to their publications

We create information about each research group manually to relate DCS members with their research groups:

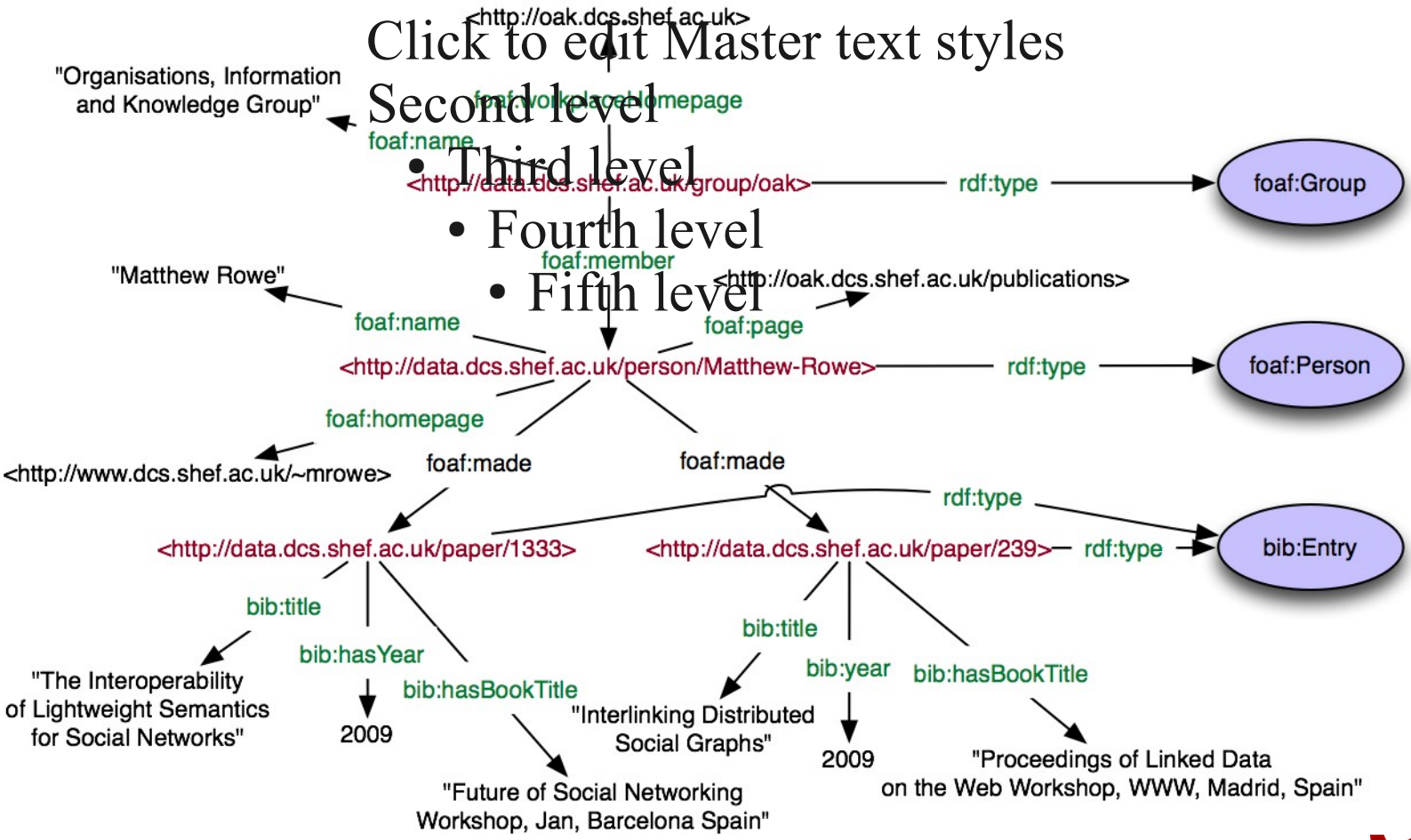
```
<http://data.dcs.shef.ac.uk/group/oak>
  rdf:type  foaf:Group ;
  foaf:name  "Organisations, Information and
Knowledge Group" ;
```


Organisations, Information and Knowledge

Coreference Resolution

Click to edit Master text styles

- Second level
  - Third level
    - Fourth level
      - Fifth level



Linking to the Web of Linked Data

Our dataset at this stage in the approach is not linked data

All links are *internal* to the dataset

To overcome the burden of researchers updating their publications we query the DBLP linked dataset using a Networked Graph SPARQL query:

The query detects authored research papers in DBLP based on coauthorship with co-workers

```
CONSTRUCT {
  ?q foaf:made ?paper .
  ?p foaf:made ?paper
}
WHERE {
  ?group foaf:member ?q .
  ?group foaf:member ?p .
  ?q foaf:name ?n .
  ?p foaf:name ?c .
  GRAPH <http://www4.wiwiss.fu-berlin.de/dblp/>
  {
    ?paper dc:creator ?x .
    ?x foaf:name ?n .
    ?paper dc:creator ?y .
    ?y foaf:name ?c .
  }
  FILTER (?p != ?q)
}
```

```
<http://data.dcs.shef.ac.uk/person/Fabio_Ciravegna>
  foaf:made  <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/icml/IresonCCFKL05> ;
  foaf:made  <http://www4.wiwiss.fu-berlin.de/dblp/resource/record/conf/ijcai/BrewsterCW01>
```

Deployment

Data.dcs is now up and running and can be accessed at the following URL:

http://data.dcs.shef.ac.uk (please try it!)

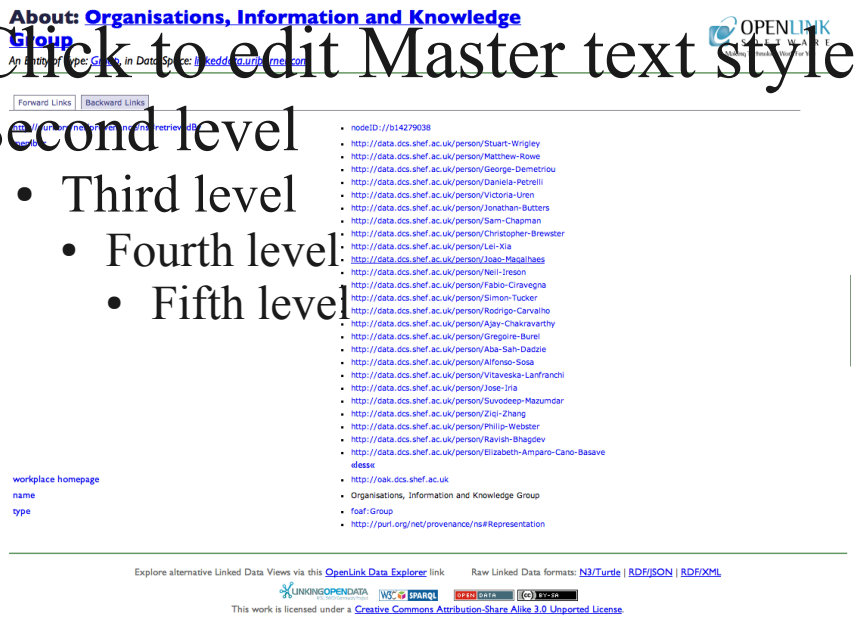
The data is deployed using

Recipe 1 from “How to Publish Linked Data”

http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Recipe 2 for Slash Namespaces from “Best Practices for Publishing RDF Vocabularies”

http://www.w3.org/TR/swbp-vocab-pub/



Viewing <http://data.dcs.shef.ac.uk/group/oak> using OpenLinks’s URIBurner

Organisations, Information and Knowledge

Conclusions

Leveraging legacy data requires information extraction components able to handle heterogeneous formats

Hidden Markov Models provide a single solution to this problem, however other methods exist which could be explored

Presented methods are applicable to other domains, simply requires a different topology and training

Current methods for Linked DCS Data into the Web of Linked Data are conservative: Bespoke SPARQL queries

Future work will include the exploration of machine learning classification techniques to perform URI disambiguation

This work is now being used as a blueprint for producing

Twitter: @mattroweshow
Web: http://www.dcs.shef.ac.uk/~mrowe
Email: m.rowe@dcs.shef.ac.uk

Questions?

(Mika et al, 2009) - Peter Mika, Edgar Meij, and Hugo Zaragoza. Investigating the semantic gap through query log analysis. In 8th International Semantic Web Conference (ISWC2009), October 2009.