Large Scale Semantic Annotation, Indexing, and Search at The National Archives

Diana Maynard

Mark Greenwood

University of Sheffield, UK

Burning questions you may have....

- In the last 3 years, which female politicians have talked about hospitals spending more than 1 million pounds?
- Which cabinet ministers currently in power were born in Sheffield?
- Which schools spent between 1 and 10 million pounds in the period July 2010-July 2011?
- Which speeches did Tony Blair make about foot and mouth disease while Prime Minister, and when did he make them?`

Government Web Archive Project

- Project aims to help open up TNA's records of .gov.uk websites (going back to 1997 and comprising some 340 million pages)
- Government funding has been allocated to publishing more and more material on data.gov.uk in open and accessible forms
- But it's still pretty hard to find the information you're looking for
- Aim is basically to improve access to this enormous volume of data - both for the general public and for specialist researchers

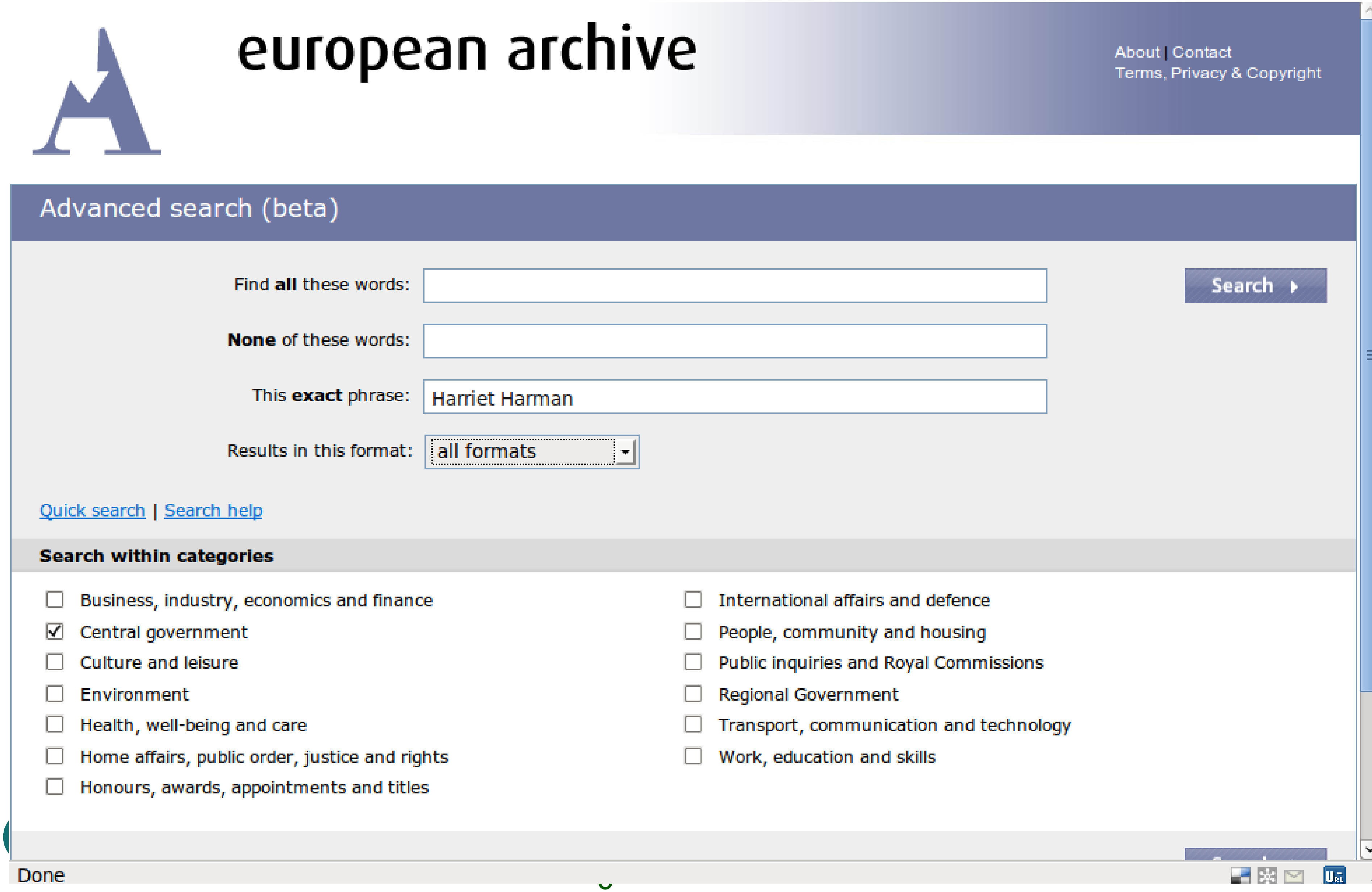
What are the National Archives?

- UK government's official archive, containing over 1,000 years of history and making this information publicly available
- Work with 250 government and public sector bodies, helping them to manage and use information more effectively.
- Over 11 million historical government and public records - one of the largest in the world.
- In general, government records that have been selected for permanent preservation are sent to The National Archives when they are 30 years old, but many are released earlier under the Freedom of Information Act

What kinds of information do they hold?

- **Government, diplomacy and the armed forces**: e.g. documents from all government departments
- **Court records**
- Approx 6 million **historical maps**
- **DocumentsOnline**: digitised public records, e.g. famous historical wills, selected records from MI5 and MI6, a range of UFO-related files from the MoD, WW1 and WW2 selected records
- Census records, alien arrivals, birth, marriage and death certificates etc.

Basic National Archives search tool



european archive

About | Contact
Terms, Privacy & Copyright

Advanced search (beta)

Find **all** these words:

None of these words:

This **exact** phrase: Harriet Harman

Results in this format: all formats

Quick search | Search help

Search within categories

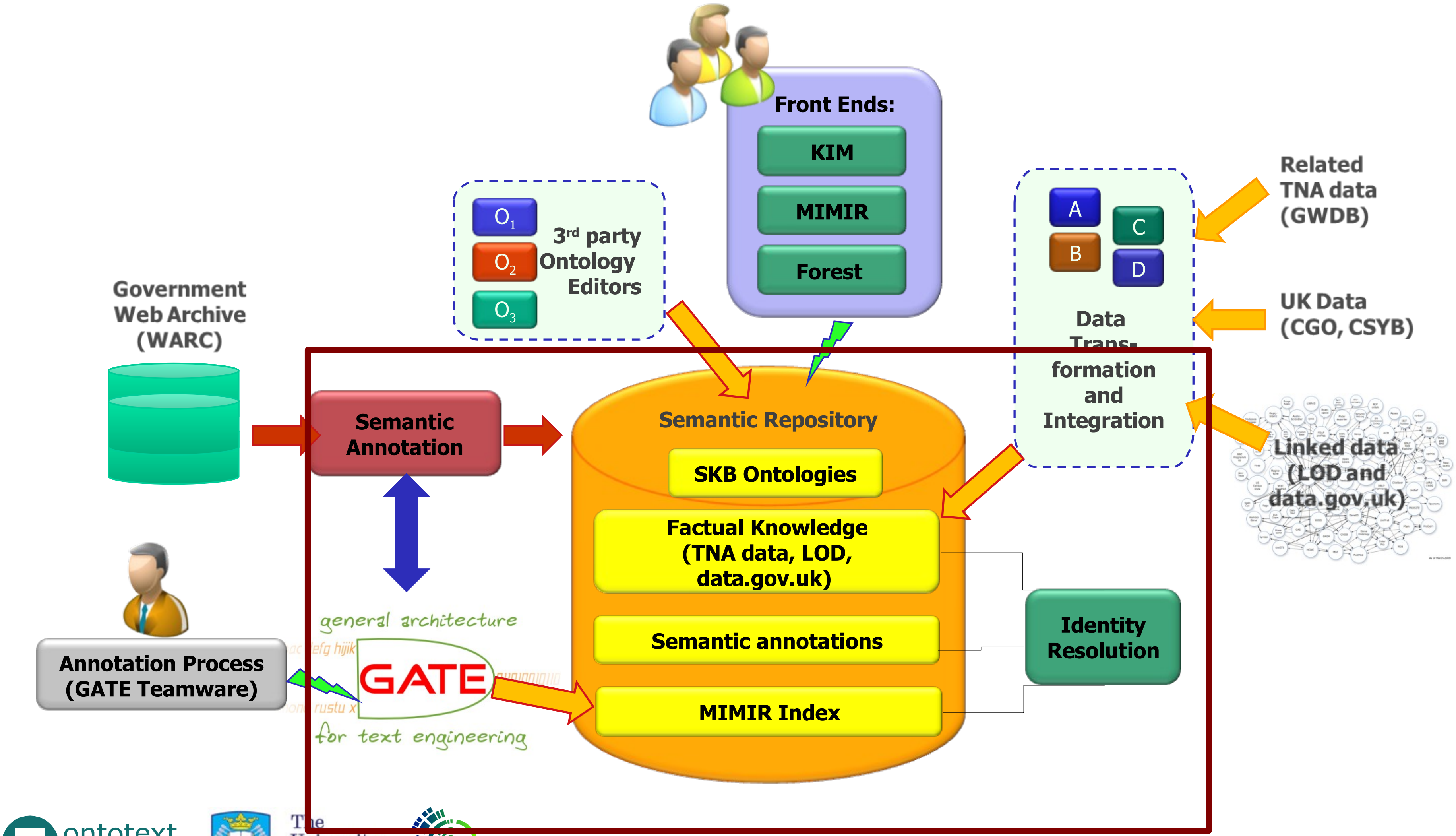
- □ Business, industry, economics and finance
- ☑ Central government
- □ Culture and leisure
- □ Environment
- □ Health, well-being and care
- □ Home affairs, public order, justice and rights
- □ Honours, awards, appointments and titles
- □ International affairs and defence
- □ People, community and housing
- □ Public inquiries and Royal Commissions
- □ Regional Government
- □ Transport, communication and technology
- □ Work, education and skills

Done

We want to do better than this!

- Enable semantic-based search for categories of things, e.g. all Cabinet Ministers, all cities in the UK
- Search results include morphological variants of words and synonyms
- Search for specific phrases with some unknowns, e.g. a Person and a monetary amount in the same sentence
- Search for ranges, e.g. all monetary amounts greater than a million pounds
- Restrict search to certain date periods, domains etc.
- and so on...

System Architecture: GATE/MIMIR



Government Web Archive (WARC)

Semantic Annotation

Annotation Process (GATE Teamware)

general architecture
GATE
for text engineering

O$_1$
O$_2$
O$_3$
3rd party Ontology Editors

Front Ends:
KIM
MIMIR
Forest

Semantic Repository
SKB Ontologies
Factual Knowledge (TNA data, LOD, data.gov.uk)
Semantic annotations
MIMIR Index

A
B
C
D
Data Transformation and Integration

Related TNA data (GWDB)

UK Data (CGO, CSYB)

Linked data (LOD and data.gov.uk)

Identity Resolution

ontotext
Semantic Technology Lab

The University Of Sheffield.

How does it work?

- Step 1: Annotate the documents using GATE
  - Ontology-based information extraction (OBIE) tools to annotate entities in text and relates them to ontology where appropriate
- Step 2: Index the documents using GATE/MIMIR
  - Create a MIMIR index based on the annotations produced in Step 1
- Step 3: Search the documents using MIMIR
  - Query the MIMIR index using full text or annotation-based queries
- Step 4: Browse the results
  - Search results point back to the texts in the original archive

Annotation Types

General NEs:

- standard ANNIE NEs (with some additional features)

Measurements:

- measurement (dimension, type, unit, value, normalised, scalar, interval)
- ratio (value)

Posts:

- cabinet, civil service, military, medical, other (e.g. MP, CEO)

Official Documents

- legislation, other (e.g. white paper)

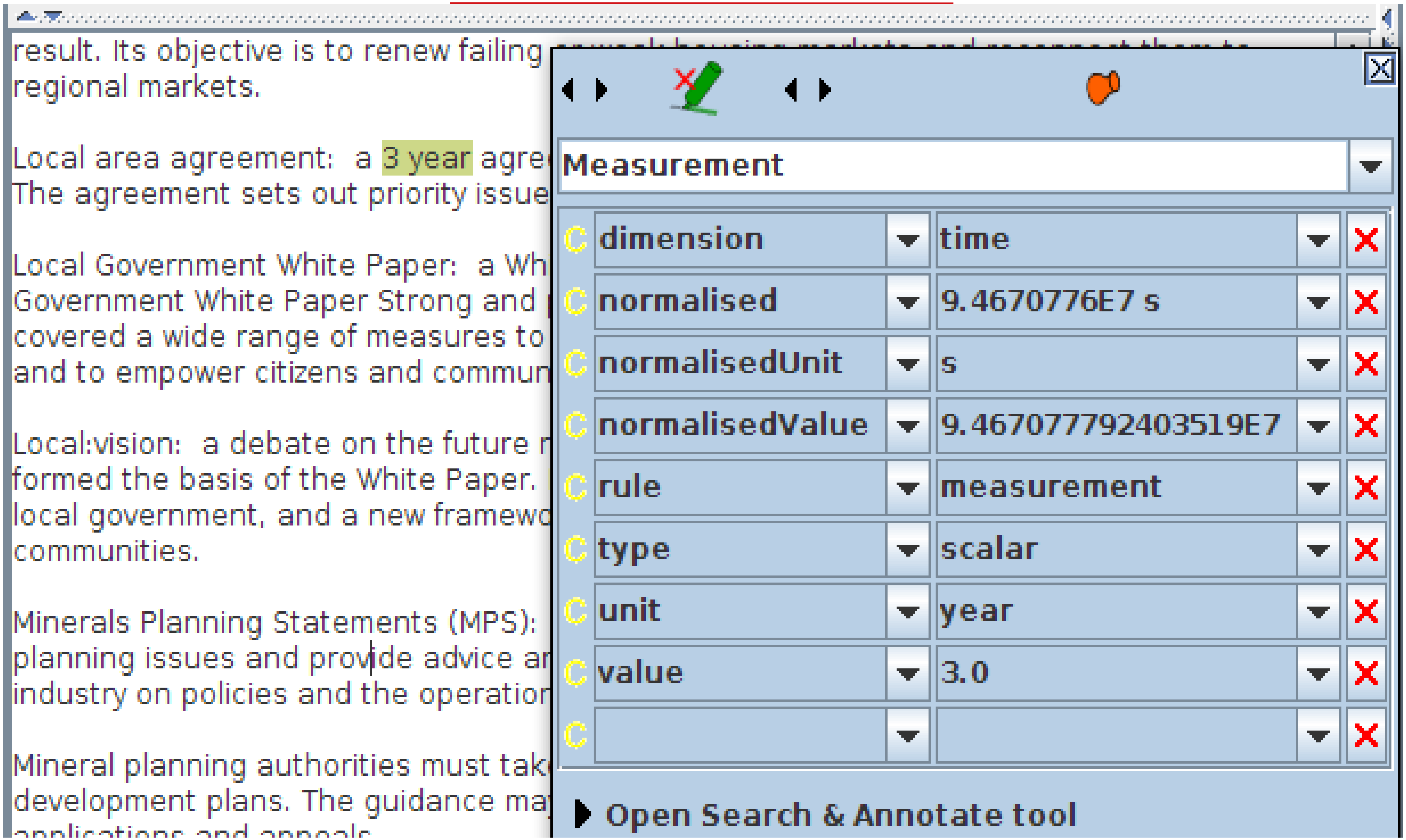
Projects/Initiatives/Campaigns

Wars/Military Conflicts

Measurements

- Measurement plugin based on the GNU units application
- Recognise numbers in documents, and then use the parser to determine if the following words are valid units
- Number and units are combined to give a measurement
- Each measurement is then normalised to SI units, e.g. all time measurements are normalised to seconds, distances to meters etc.
- This enables searching for measurements no matter how they are expressed in the text, e.g. searching for 0.03m would also find references to 3cm, 30mm and 1.18 inches.

Measurement annotation



result. Its objective is to renew failing
regional markets.

Local area agreement: a 3 year agre
The agreement sets out priority issue

Local Government White Paper: a Wh
Government White Paper Strong and
covered a wide range of measures to
and to empower citizens and commun

Local:vision: a debate on the future
formed the basis of the White Paper.
local government, and a new framewo
communities.

Minerals Planning Statements (MPS):
planning issues and provide advice ar
industry on policies and the operatior

Mineral planning authorities must tak
development plans. The guidance ma
applications and appeals

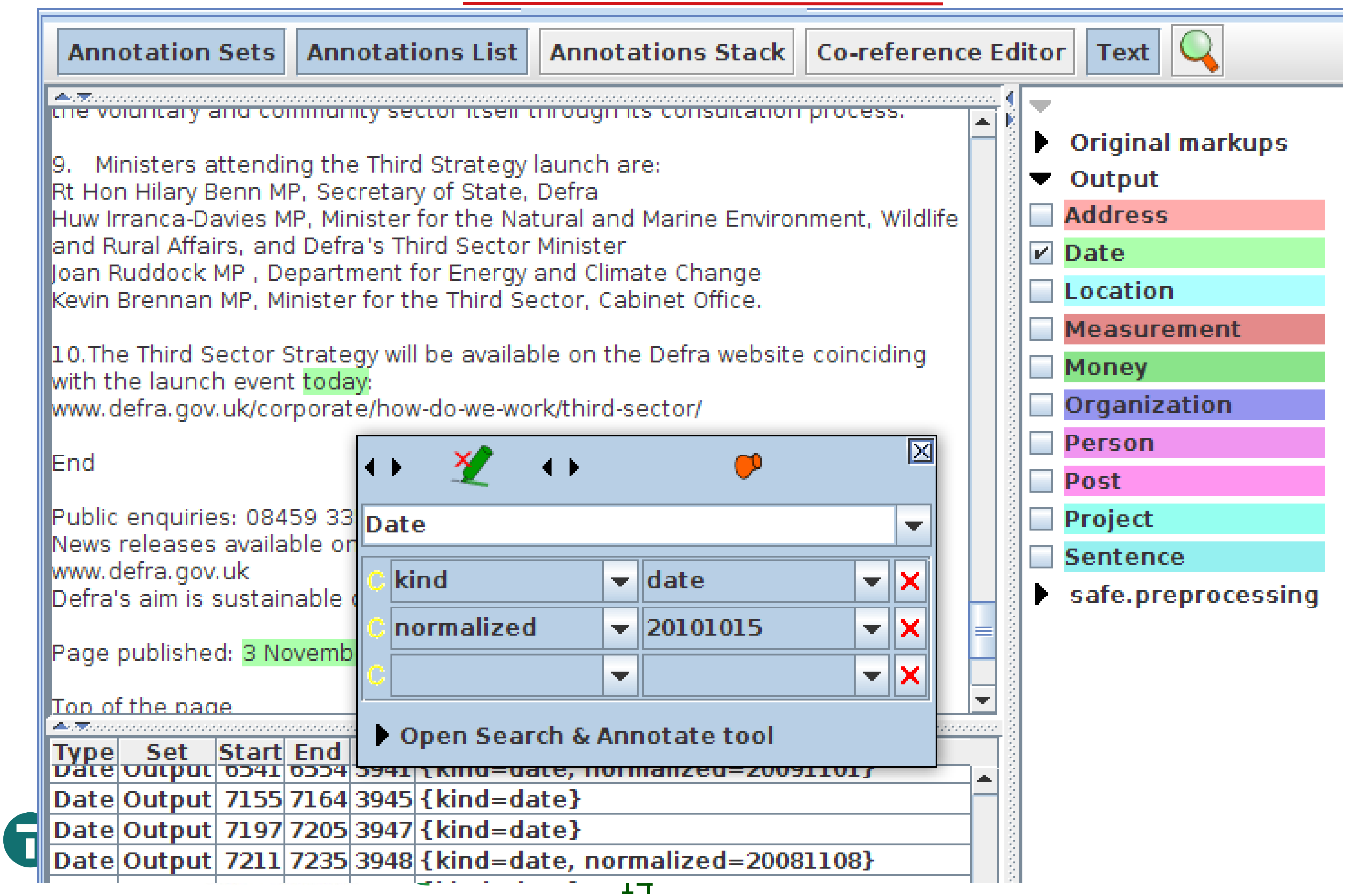
Measurement

dimension	time
normalised	9.4670776E7 s
normalisedUnit	s
normalisedValue	9.467077792403519E7
rule	measurement
type	scalar
unit	year
value	3.0

Relative date normalisation

- Absolute dates have an annotation feature showing the normalised form
- Special plugin that converts the fully specified date into standard numerical format
- If we know the date of the article, we can also calculate the actual value of relative dates (e.g. “last year”, “in the next fortnight”, etc.)
- We can often find the date of the article from the title or other information in the body of the article: this is added as a document feature
- The default option is to use the date the document was crawled
- Once we have a reference date for the document, we can normalise the values against it

Relative Date annotation



the voluntary and community sector itself through its consultation process.

9. Ministers attending the Third Strategy launch are:
Rt Hon Hilary Benn MP, Secretary of State, Defra
Huw Irranca-Davies MP, Minister for the Natural and Marine Environment, Wildlife and Rural Affairs, and Defra's Third Sector Minister
Joan Ruddock MP , Department for Energy and Climate Change
Kevin Brennan MP, Minister for the Third Sector, Cabinet Office.

10.The Third Sector Strategy will be available on the Defra website coinciding with the launch event today:
www.defra.gov.uk/corporate/how-do-we-work/third-sector/

End

Public enquiries: 08459 33
News releases available or
www.defra.gov.uk
Defra's aim is sustainable

Page published: 3 Novemb

Top of the page

Date

kind	date
normalized	20101015

▶ Original markups
▼ Output
□ Address
☑ Date
□ Location
□ Measurement
□ Money
□ Organization
□ Person
□ Post
□ Project
□ Sentence
▶ safe.preprocessing

Type	Set	Start	End		
Date	Output	6541	6554	3941	{kind=date, normalized=20091101}
Date	Output	7155	7164	3945	{kind=date}
Date	Output	7197	7205	3947	{kind=date}
Date	Output	7211	7235	3948	{kind=date, normalized=20081108}

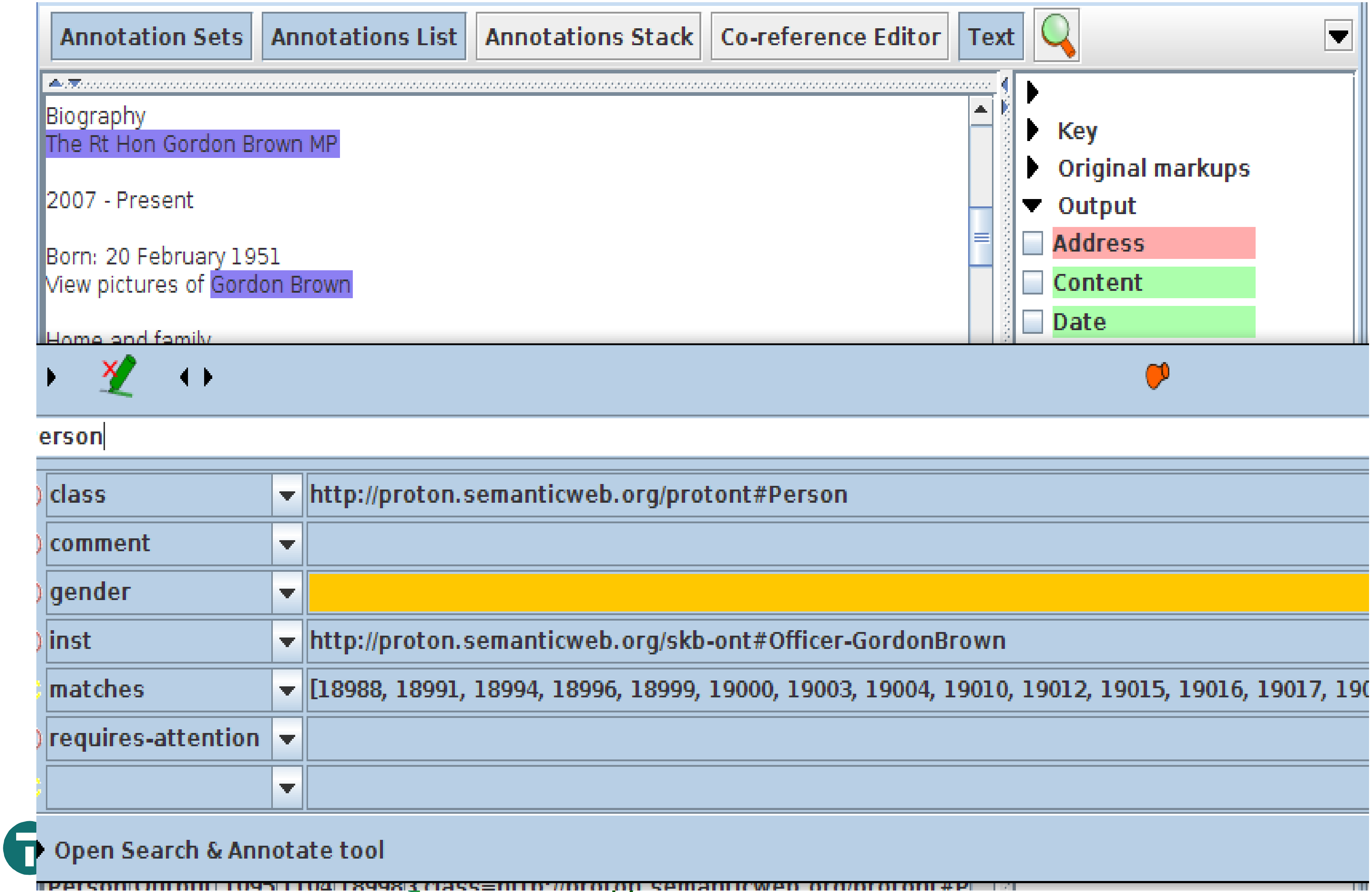
Semantic Knowledge Base (SKB)

- Semantic Knowledge Base developed (by Ontotext) on the basis of information extracted from the archive, along with existing KBs and ontologies
- Pre-processed, transformed and integrated to form a consistent KB
- Continuously enriched and extended
- Contains linked data from LOD cloud, data from data.gov.uk, data from related TNA projects, geographical data from Ontotext
- We don't try to extract, for example, a consistent description of London from the archive
  - Instead we integrate various knowledgebases and ontologies where extensive and well-maintained geographical features are already present
  - So a reference to London from a web page in the archive will be annotated with reference to the profile of London in the background knowledge bases

Using SKB to Annotate the Archive

- Standard IE uses small hand-crafted gazetteers to annotate relevant government entities.
- Semantic Knowledge Base developed (by Ontotext) on the basis of information extracted from the archive, along with existing KBs and ontologies
- The SKB contains a lot more relevant information useful for annotation
- Use the LKB gazetteer to annotate certain concepts directly from the ontology
  - his leads to greater coverage and (theoretically) better precision
  - Where appropriate, annotated entities are linked to specific instances in the ontology, for example Post
- Identity resolution module also links co-referring mentions with a known URI

Co-referring mentions linked to instance



Biography
The Rt Hon Gordon Brown MP

2007 - Present

Born: 20 February 1951
View pictures of Gordon Brown

Home and family

Key
Original markups
Output
□ Address
□ Content
□ Date

erson

class	http://proton.semanticweb.org/protont#Person
comment	
gender	
inst	http://proton.semanticweb.org/skb-ont#Officer-GordonBrown
matches	[18988, 18991, 18994, 18996, 18999, 19000, 19003, 19004, 19010, 19012, 19015, 19016, 19017, 190
requires-attention	

GATE Cloud Paralleliser

- One of the main challenges in the task of annotating, indexind and searching millions of documents is the sheer size of the data
- At the time of processing, the archive contained 42TB of data (around 700 million documents, of which around 150 million were unique)
- We processed the data using GATE Cloud Parallelizer, installed on the Amazon Cloud https://gatecloud.net/
- GCP is a platform for parallel semantic annotation of documents, designed as a parallel version of the execution engine installed in GATE
- It takes a language processing pipeline created using GATE Developer, and executes it using a series of parallel threads
- The job control is performed by executing batches, which are XML files describing outstanding tasks

GATE Mímir: Answering Questions Google Can't

Mímir

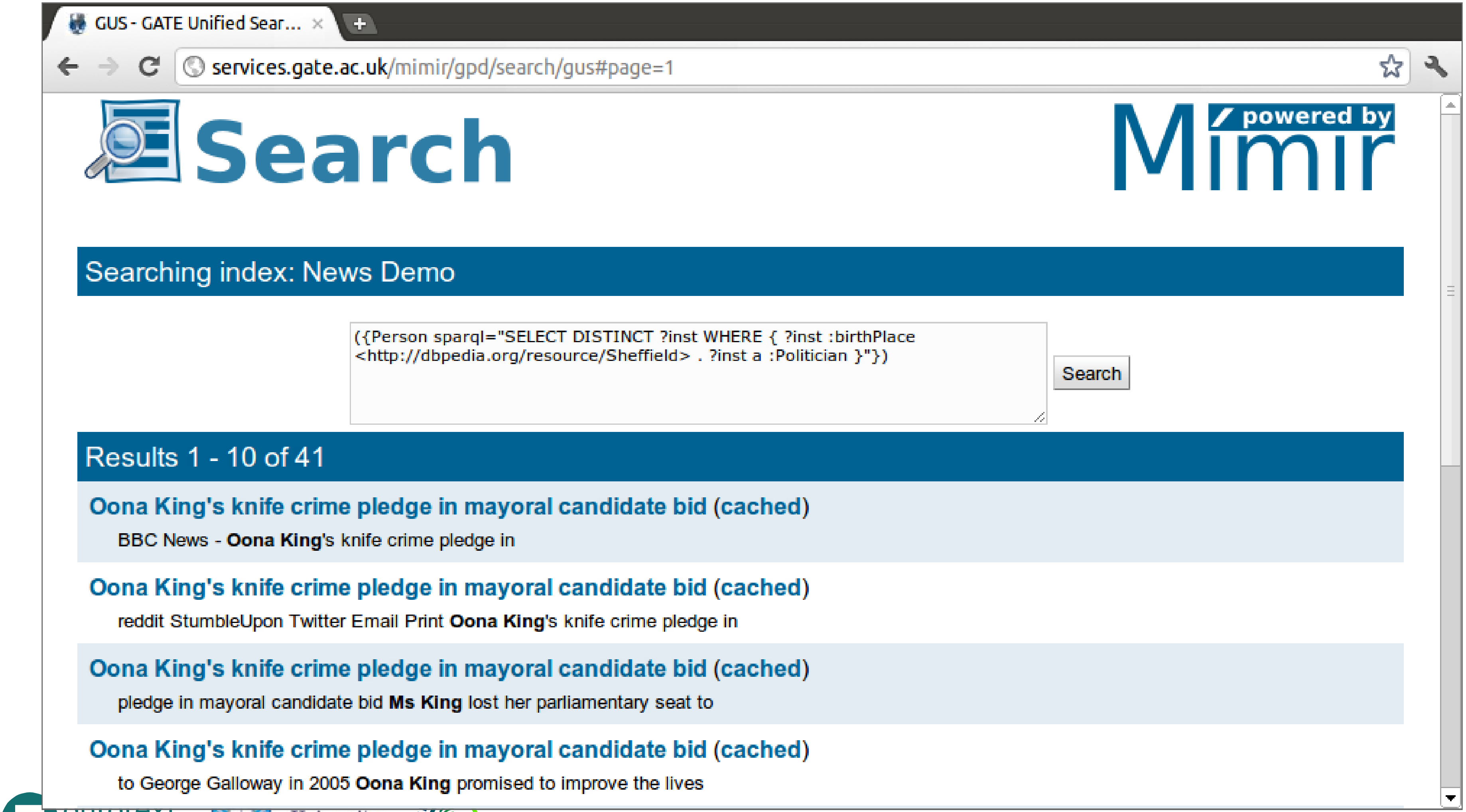
- Mímir is an IR engine that can index and search over:
  - text
  - semantic annotations
  - ontologies and KBs
- Allows queries that arbitrarily mix full-text, structural, semantic and linguistic annotations
- Scales to millions of documents

What can GATE Mímir do that Google can't?

Show me:

- all documents mentioning a temperature between 30 and 90 degrees F (expressed in any unit)
- all abstracts written in French on Patent Documents from the last 6 months which mention any form of the word “transistor” in the English abstract
- the names of the patent inventors of those abstracts
- all documents mentioning steel industries in the UK, along with their location

Search News Articles for Politicians born in Sheffield



GUS - GATE Unified Sear...
services.gate.ac.uk/mimir/gpd/search/gus#page=1
Search
powered by Mímir

Searching index: News Demo

```
({Person sparql="SELECT DISTINCT ?inst WHERE { ?inst :birthPlace
<http://dbpedia.org/resource/Sheffield> . ?inst a :Politician }"})
```

Results 1 - 10 of 41

Oona King's knife crime pledge in mayoral candidate bid (cached)

BBC News - **Oona King**'s knife crime pledge in

Oona King's knife crime pledge in mayoral candidate bid (cached)

reddit StumbleUpon Twitter Email Print **Oona King**'s knife crime pledge in

Oona King's knife crime pledge in mayoral candidate bid (cached)

pledge in mayoral candidate bid **Ms King** lost her parliamentary seat to

Oona King's knife crime pledge in mayoral candidate bid (cached)

to George Galloway in 2005 **Oona King** promised to improve the lives

http://demos.gate.ac.uk/mimir/gpd/search/gus

Summary

- Use of simple mechanisms can add value and increase usage in the short and medium terms to a huge archive of information.
- Our GATE-based text-mining solution provides search paradigms over semantic annotation that relates archival content to Linked Data and other structured sources.
- Using a Semantic KB makes it possible to integrate external knowledge with facts extracted from texts
- Initial evaluations by TNA (comparing it with their existing search tools) were very promising
- Tools will continue to be developed and enhanced by in-house TNA team, and new interfaces developed