Making the UK Government Web Archive More Usable

Tom Storrar

21 June 2012

The National Archives

- The UK Government Web Archive and Web Continuity
- Who uses the UK Government Web Archive, and what for?
- Web Archives: The problem of discovery
- Why a semantic knowledge base?
- How did we go about it?
- Putting the knowledge base live

The UK Government Web Archive and Web Continuity

What is Web Archiving?

Web archiving is the process of collecting portions of the World Wide Web and ensuring the collection is preserved in an archive, such as an archive site, for future researchers, historians, and the public. Due to the massive size of the Web, web archivists typically employ web crawlers for automated collection.

(http://en.wikipedia.org/wiki/Web_archiving, 01/07/2012)

The UK Government Web Archive

- The Internet Memory Foundation have been contracted to crawl and host the collection since 2005.
- Is free to use and is accessible online.
- Contains approximately 1 billion documents.
- There is a great variety of content, archived over many years.
- Has a wide variety of users

Web archive examples

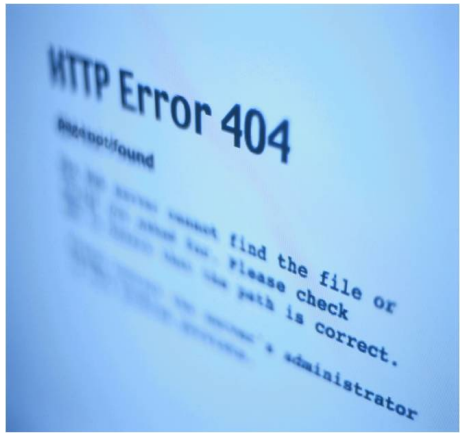
The National Archives

The UK Government and the Web

- Massive growth in use of the web by UK Central Government.
- We need to engage or face losing significant parts of the evidence base.
- The nature of the web poses unique challenges and unique opportunities.
- Difficulty of distinction between records and information.

Web Continuity

- Comprehensively web archiving the UK central government web estate.
- Supports the Website Review programme.
- Redirection solution.

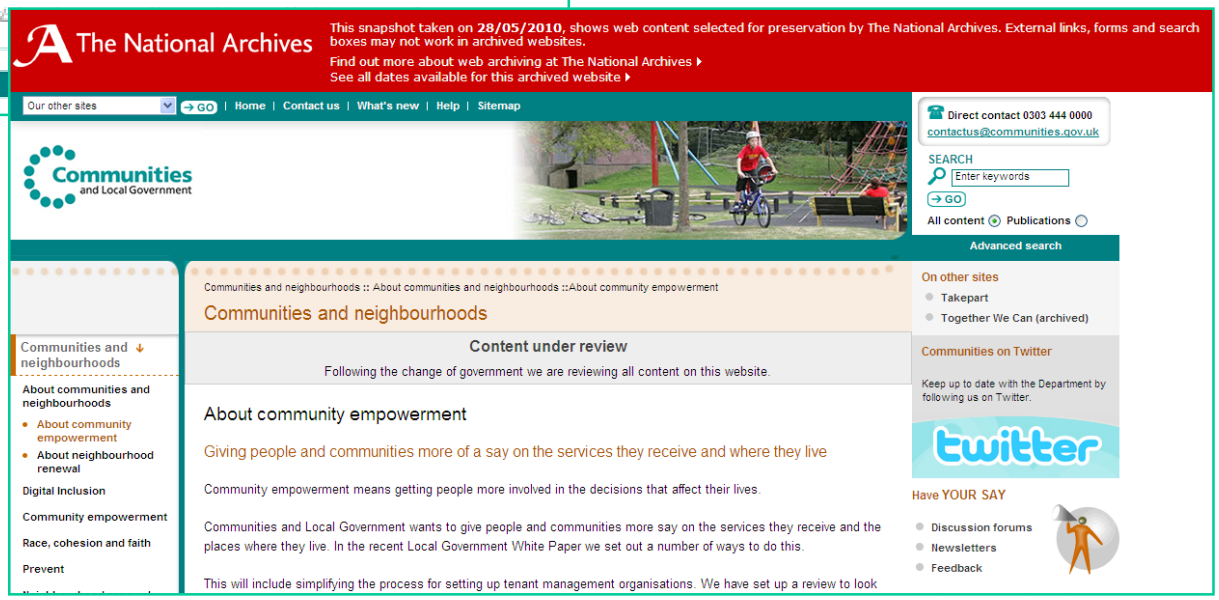


- URLs in Hansard and official publications can work in perpetuity.

From this:

Redirection

To this:



The 2010 General Election

- We anticipated that content would be removed from websites and we saw this happen.
- Government departments refer users to the web archive either by redirection or by linking.
- Increased usage of the web archive.
- Many other examples of thematic content – for example, London 2012.

UK Government Web Archive

http://www.nationalarchives.gov.uk/webarchive/

2010 General Election: Before & After

http://webarchive.nationalarchives.gov.uk/20100505180303/http://www.number10.gov.uk/

http://webarchive.nationalarchives.gov.uk/20100512155632/http://www.number10.gov.uk/

Who are our users and what do they want?

Who are our users?

- We conducted user surveys in May 2010 and August 2011
- A great variety of users
- Existing users and potential users

What do our users want?

The three areas of the web archive that are consistently most popular amongst respondents were:

- **Health, well-being and care**
- **Work, education and skills**
- **Home affairs, public order, justice and rights.**

The most needed improvement suggested was to search functionality

Sketching out User Stories

User stories based on survey results, experience of web archiving team and understanding of potential of technology:

“As a member of the public I'm trying to find out which department is responsible for what BERR used to be responsible for and access the old content”

“As a researcher I want to find all guidance related to wind turbines published between 2005 and 2008 on websites in the "Environment" category”

Web Archives: The problem of discovery

The problems with conventional search

- **Scale** - The UKGWA is very large! Around 1B pages and 80TB of data.

- **Duplication -** Information, data and documents may exist in the web archive in multiple instances.

- **Structure -** Government websites are all structured in different ways and the names and functions of departments change from time to time

- **Domain knowledge** - Searching for specific subjects requires the enquirer to be aware of which department was responsible for particular functions over time

Why a semantic knowledge base?

• **Scale** - The size of the web archive means that we needed a more efficient and intuitive approach

• **Duplication –** Can be solved by identifying duplicates and skipping them being processed

• **Structure –** Co-referencing and use of ontologies provides more consistency and context.

• **Domain knowledge** – Users can benefit from explicit knowledge without knowing

• **Linked data –** Allows the system to interact with linked open data on the web

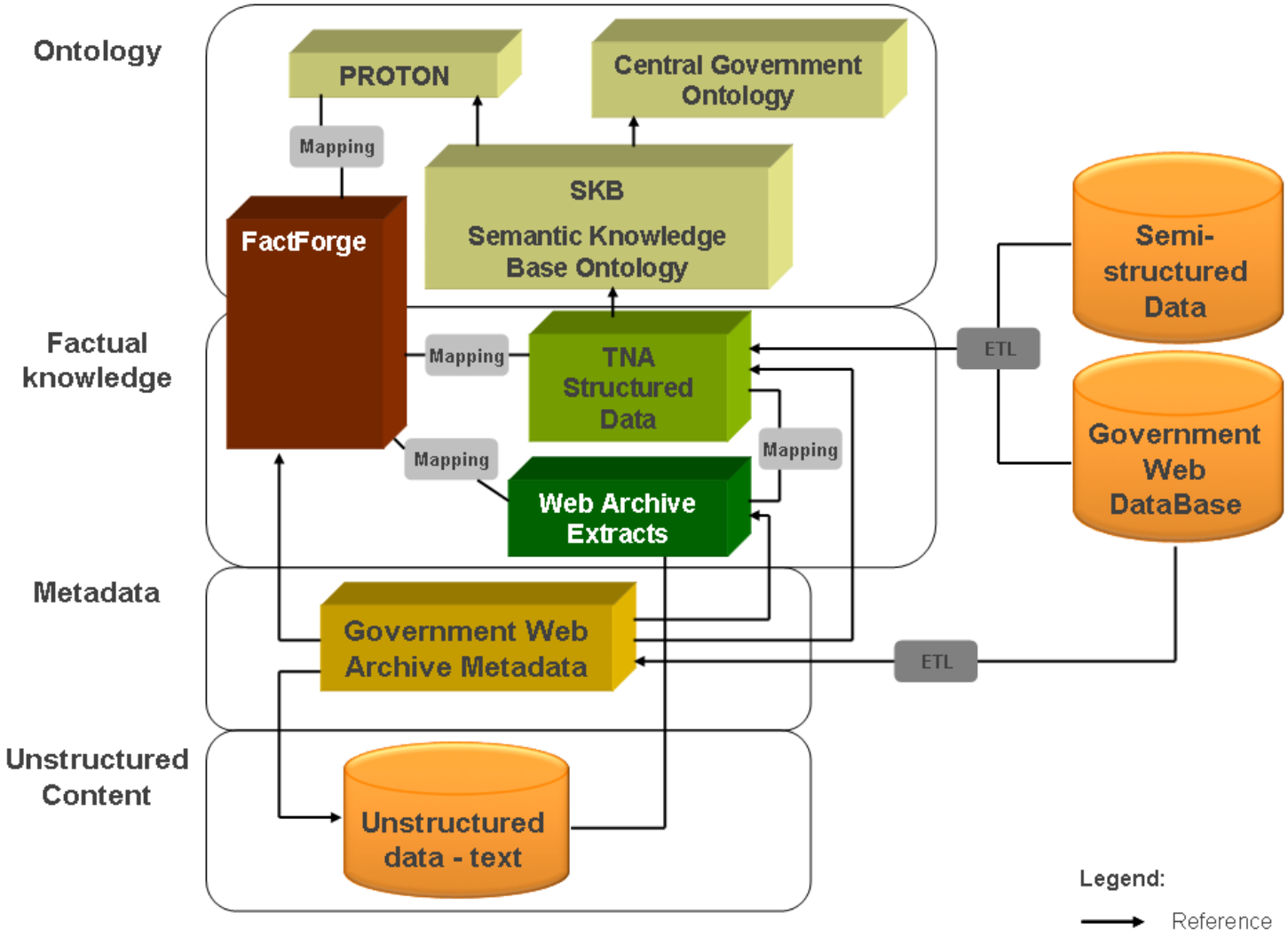
How did we go about it?

The Semantic Knowledge Base – Products

The SKB is built using component products supplied by GATE Team and Ontotext:

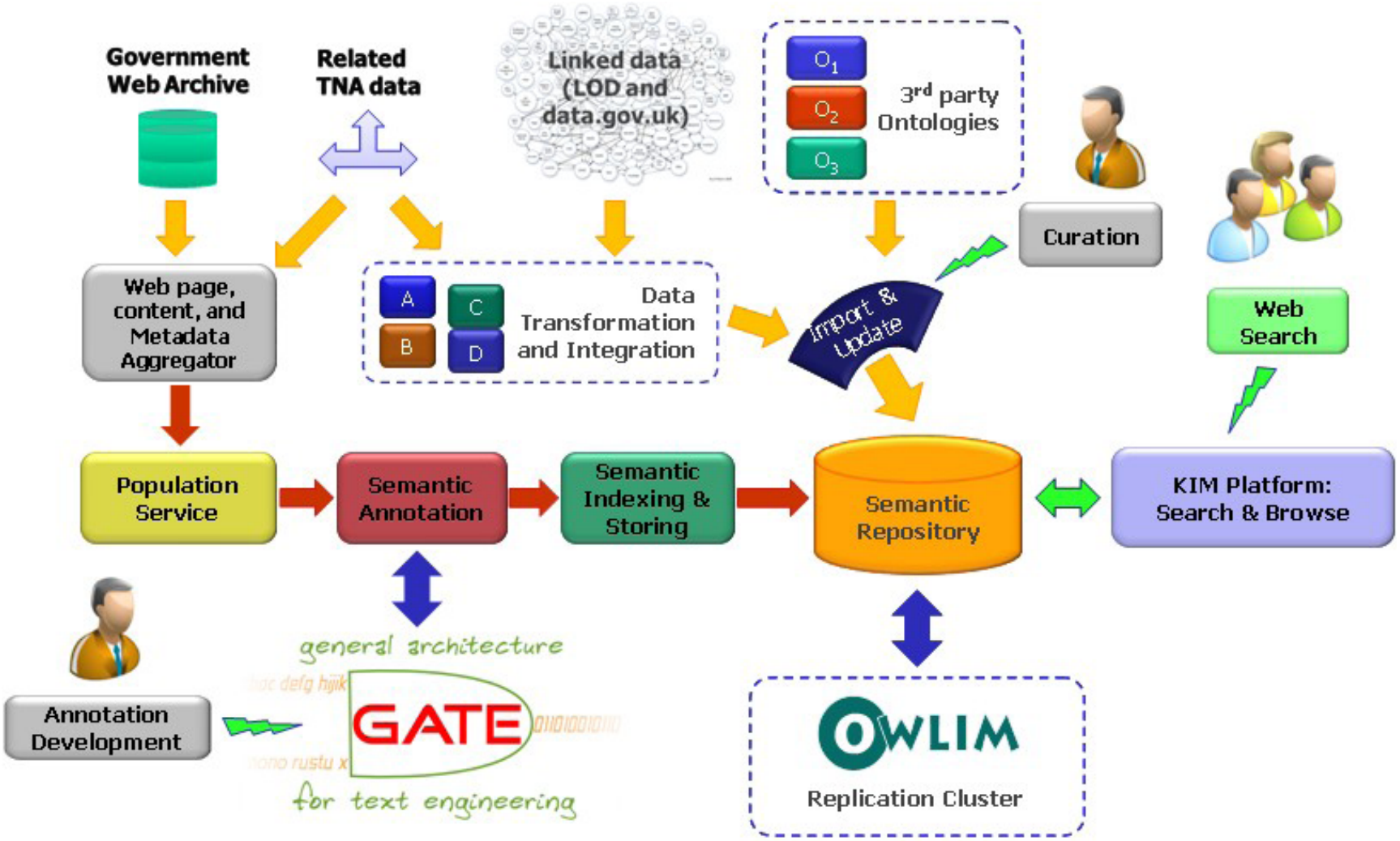
- GATE Teamware
- GATE Developer
- GATE Mimir/GUS
- Ontotext FactForge
- Ontotext BigOWLIM
- Ontotext KIM
- Ontotext Forest

Conceptual Architecture



The National Archives

Main System Components



The National Archives

Scope of the knowledge base

• All major central government departments' and some other bodies' websites selected

• Web archive content from 1997 to end of 2011

• Resulting in a subset of about 35 million documents to be processed by GATE

What annotations to make?

Decisions based on:

- User surveys
- Knowledge within The National Archives
- The nature of the web archive (specific domain but very heterogeneous within it!)
- Practical queries that will help our users.

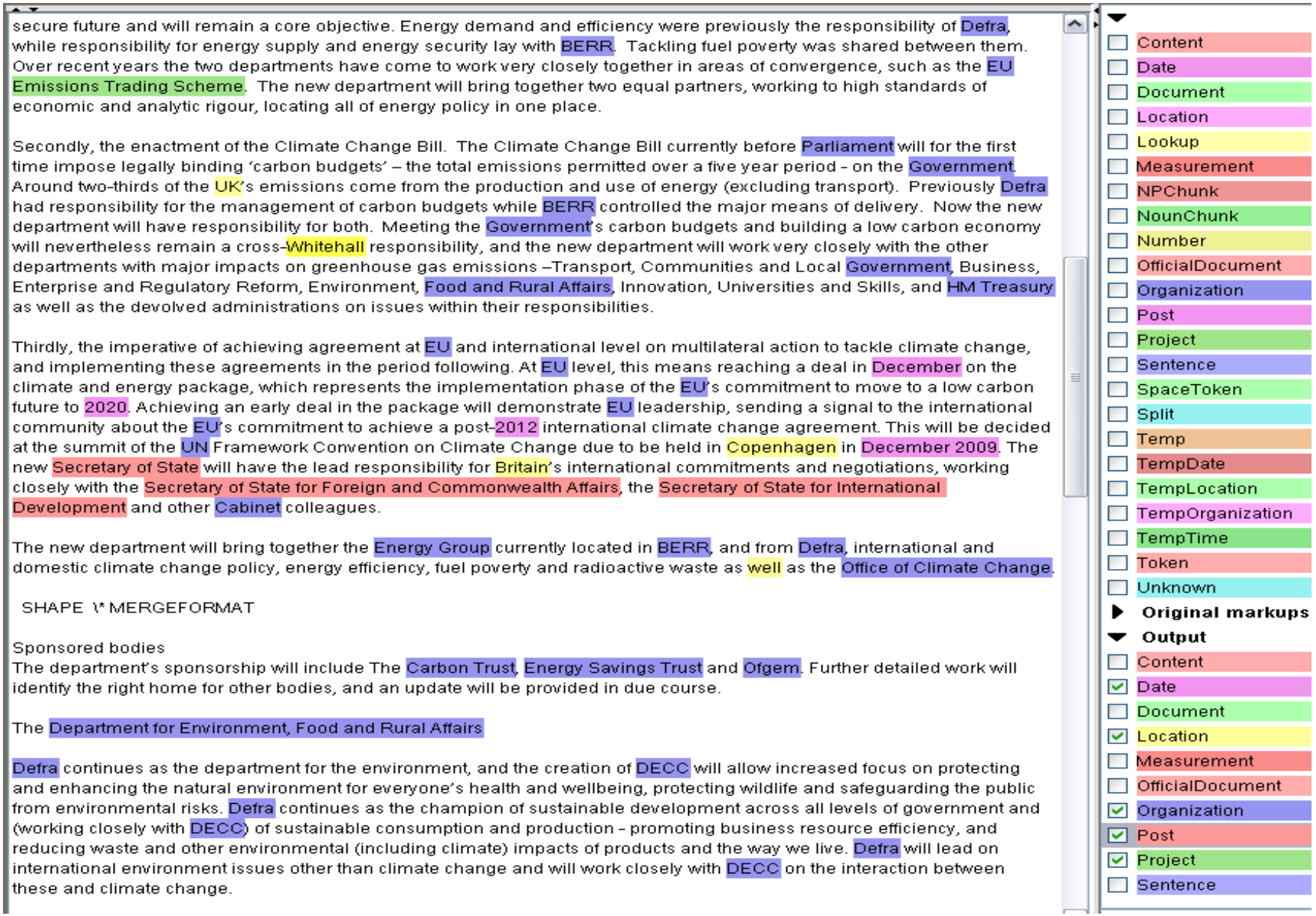
Annotation types used

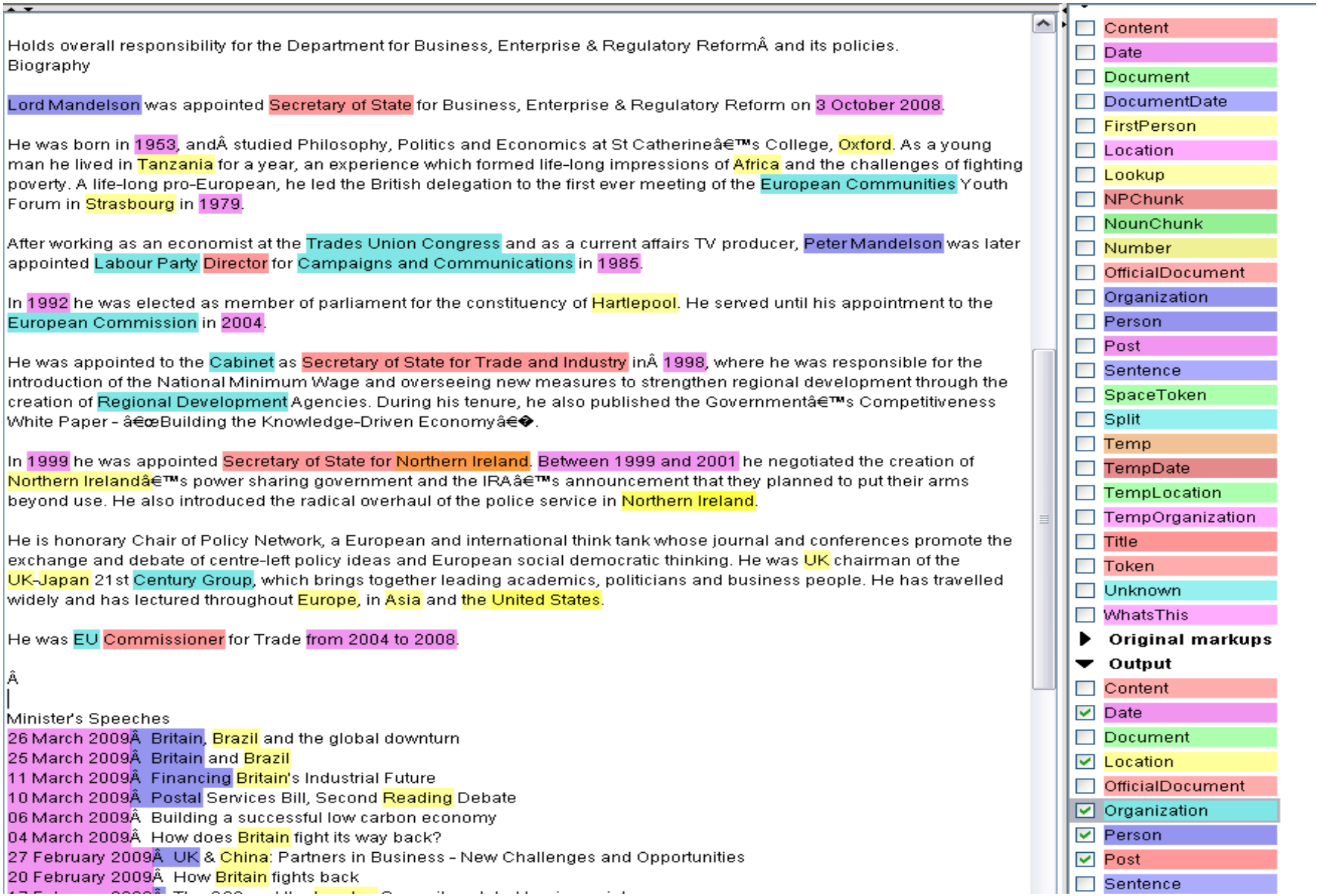
Annotations picked up with ANNIE, with some modification:

Content
Person
Location - modified and added some sub-types (constituency, county etc)
Date
Money
Address
Etc…

Customised and modified types and sub-types

Annotation	Use	Annotation type/sub-type	Example annotation
Post	To identify positions within government	Post kind=cabinet	[Secretary of State for Work and Pensions]
		Post kind=civil service	[Permanent Secretary]
Organization	To identify specific government organisations	Organization orgType=ministerial department	[Home Office]
		Organization orgType=agency	[Criminal Records Bureau] [DVLA]
Conflict	To annotate any major military conflict or war	Conflict	[Vietnam War] [Falklands Conflict]
Military Operation	Used to annotate any military operation	MilitaryOperation	[Operation Grand Slam] [Operation Reindeer]
Project	To annotate any project or initiative	Project	[NHS Health Trainers Initiative] [FIFA World Cup 1GOAL campaign]


secure future and will remain a core objective. Energy demand and efficiency were previously the responsibility of Defra, while responsibility for energy supply and energy security lay with BERR. Tackling fuel poverty was shared between them. Over recent years the two departments have come to work very closely together in areas of convergence, such as the EU Emissions Trading Scheme. The new department will bring together two equal partners, working to high standards of economic and analytic rigour, locating all of energy policy in one place.
Secondly, the enactment of the Climate Change Bill. The Climate Change Bill currently before Parliament will for the first time impose legally binding 'carbon budgets' – the total emissions permitted over a five year period - on the Government. Around two-thirds of the UK's emissions come from the production and use of energy (excluding transport). Previously Defra had responsibility for the management of carbon budgets while BERR controlled the major means of delivery. Now the new department will have responsibility for both. Meeting the Government's carbon budgets and building a low carbon economy will nevertheless remain a cross-Whitehall responsibility, and the new department will work very closely with the other departments with major impacts on greenhouse gas emissions –Transport, Communities and Local Government, Business, Enterprise and Regulatory Reform, Environment, Food and Rural Affairs, Innovation, Universities and Skills, and HM Treasury as well as the devolved administrations on issues within their responsibilities.
Thirdly, the imperative of achieving agreement at EU and international level on multilateral action to tackle climate change, and implementing these agreements in the period following. At EU level, this means reaching a deal in December on the climate and energy package, which represents the implementation phase of the EU's commitment to move to a low carbon future to 2020. Achieving an early deal in the package will demonstrate EU leadership, sending a signal to the international community about the EU's commitment to achieve a post-2012 international climate change agreement. This will be decided at the summit of the UN Framework Convention on Climate Change due to be held in Copenhagen in December 2009. The new Secretary of State will have the lead responsibility for Britain's international commitments and negotiations, working closely with the Secretary of State for Foreign and Commonwealth Affairs, the Secretary of State for International Development and other Cabinet colleagues.
The new department will bring together the Energy Group currently located in BERR, and from Defra, international and domestic climate change policy, energy efficiency, fuel poverty and radioactive waste as well as the Office of Climate Change.
SHAPE * MERGEFORMAT
Sponsored bodies
The department's sponsorship will include The Carbon Trust, Energy Savings Trust and Ofgem. Further detailed work will identify the right home for other bodies, and an update will be provided in due course.
The Department for Environment, Food and Rural Affairs
Defra continues as the department for the environment, and the creation of DECC will allow increased focus on protecting and enhancing the natural environment for everyone's health and wellbeing, protecting wildlife and safeguarding the public from environmental risks. Defra continues as the champion of sustainable development across all levels of government and (working closely with DECC) of sustainable consumption and production - promoting business resource efficiency, and reducing waste and other environmental (including climate) impacts of products and the way we live. Defra will lead on international environment issues other than climate change and will work closely with DECC on the interaction between these and climate change.
Content
Date
Document
Location
Lookup
Measurement
NPChunk
NounChunk
Number
OfficialDocument
Organization
Post
Project
Sentence
SpaceToken
Split
Temp
TempDate
TempLocation
TempOrganization
TempTime
Token
Unknown
Original markups
Output
Content
Date
Document
Location
Measurement
OfficialDocument
Organization
Post
Project
Sentence



The National Archives

Annotation process

- Classic, iterative GATE process across carefully-selected corpus of representative documents from the web archive

- Used GATE Teamware for 3 rounds of manual annotation after the modified ANNIE ran over the corpus

- Analysis showed that inter-annotator agreement was quite high

- In collaboration with the GATE team in Sheffield

…and after GATE

- Triple store, OWLIM, holds triples
- SPARQL querying allows queries of RDF in triple store
- Sesame tool carries query
- Linked data integration
- SKB will be updated on a quarterly basis

Example 1

"Please give me all people who have "government" in their job title, who were in that post in 2007 or since and have "john" in their name"

```
{Person semanticConstraint="
  ?inst <http://proton.semanticweb.org/skb-ont#hasPosition> ?position .
  ?position <http://proton.semanticweb.org/skb-ont#withinOrganization> ?organization .
  ?organization <http://factforge.net/preferredLabel> ?label .
  FILTER (regex(?label, 'government', 'i')) .
  ?position <http://proton.semanticweb.org/skb-ont#heldFrom> ?dateFrom .
  FILTER (str(?dateFrom) >= \"2007\") .
"} OVER john
```

Example 2

"Please give me all mentions of a person in close proximity to a date on the Cabinet Office website "

{Person} [0..5] (2007 IN {Date}) IN ({Content} IN {Document domain="www.cabinetoffice.gov.uk"})

Example 3

"Please give me all Ministers of State who have made a statement or speech"

```
{Person
  semanticConstraint="
    ?inst <http://proton.semanticweb.org/skb-ont#hasPosition> ?pos .
    ?pos <http://proton.semanticweb.org/skb-ont#hasTitle>
      <http://proton.semanticweb.org/skb-ont#OfficialTitle-Minister_of_State> ."
} root:say
```

Example 4

“Please give me all people that hold a position within an organisation with “rural” in its name”

```
{Person semanticConstraint="
  ?inst  <http://proton.semanticweb.org/skb-ont#holdsPositionOrganization> ?posorg .
  ?posorg <http://proton.semanticweb.org/skb-ont#positionWithinOrganization> ?org .
  ?org <http://factforge.net/preferredLabel> ?label .
  FILTER regex(?label, 'rural', 'i') .
"}
```

The results

- System development complete
- Testing shows great improvement over full text search
- Wider applications identified
- Some features developed applied to new GATE releases

What we've learned from using GATE

- Explain the concept clearly to stakeholders
- Encourage a collaborative approach to annotation, with both domain experts and those with less specific knowledge
- Be prepared to tolerate some false +s and -s
- Iterate the process and gradually increase the size and complexity of the annotations, as necessary
- Make sure you develop the right skills

Putting the knowledge base live

The front end

- 2 key sets of users:
  - Non-expert users will use the system by browsing The National Archives' Discovery system
  - Expert/technical users will be able to access the Mimir and triple store APIs to access the raw data.

  The Mimir interface will be used in-house, but we will use the APIs to drive the exploration of the content for public users.

Thank you

Any questions?

Please visit www.nationalarchives.gov.uk/webarchive/ to see the web archive and for more information

webarchive@nationalarchives.gsi.gov.uk

@UkNatArchives


The National Archives

Examples

- Variety of Web Archive material
- http://webarchive.nationalarchives.gov.uk/20090204022045/http://www.english-heritage.org.uk/
- http://webarchive.nationalarchives.gov.uk/20090805045007/http://www.opsi.gov.uk/
- http://web.archive.org/web/19970707050911/http://www.hm-treasury.gov.uk/index.html
- http://webarchive.nationalarchives.gov.uk/20100401161705/http://www.hm-treasury.gov.uk/
- http://webarchive.nationalarchives.gov.uk/20100706002452/http://www.ukbms.org/docs/reports/2004/BMSRpt0405.pdf
- http://webarchive.nationalarchives.gov.uk/20100830135257/http://www.dfid.gov.uk/Documents/publications/evaluation/ev662.pdf

- Redirection Examples
- http://www.justice.gov.uk/publications/consultation-pandemic-flu.htm
- http://webarchive.nationalarchives.gov.uk/+/http://www.dti.gov.uk/about/dti-ministerial-team/page8414.html
- http://www.nationalarchives.gov.uk/news/stories/9.htm

- General Election
- http://webarchive.nationalarchives.gov.uk/20100505180303/http://www.number10.gov.uk/
- http://webarchive.nationalarchives.gov.uk/20100512155632/http://www.number10.gov.uk/

- Datasets
- http://webarchive.nationalarchives.gov.uk/20100408182338/http://www.decc.gov.uk/en/content/cms/statistics/fuelpov_stats/fuelpov_stats.aspx
- http://webarchive.nationalarchives.gov.uk/*/http://rds.homeoffice.gov.uk/rds/pdfs10/hosb1010tabs.xls