GeoParsing: the digitization and historical georeferencing of text documents

Stuart Dunn
Centre for e-Research, King’s College London

ISGC, Taipei

10th March 2010
• Bicameral parliament at Stormont 1921-1972
• Transcripts of all debates - *Hansards*
• Fundamental aim - to broaden access
• 2004: Digitization of Lower House Hansards (80 volumes)

• 2008: Digitization of Upper House Hansards (53 volumes)

• Aim is to co-locate the collections in a single, sustainable repository

• Georeferencing, based on NER approach
Georeferencing: basic principles

- *Informal*: based on placenames
- *Formal*: based on coordinates, or some other mathematical expression

Benefits

- Resolving ambiguity
- Ease of access to data objects
- Integration of data from heterogeneous sources
- Resolving space and time
Mr. Fitt: It all depends what you interpret as unnecessary.

Mr. Speaker: Would hon. Members allow the business of the House to be conducted in an orderly manner?

Captain Brooke: Thank you very much, Mr. Speaker. That will make life much easier for me.

I should like to ask my right hon. and gallant Friend what measures he can take to protect our pig industry. This is something very close to my heart as we have the Ulster Swift factory in Enniskillen. Great damage is being done to us. Indeed, this smuggling is being subidised in a way by the British taxpayer. Something should be done to stop it.

The Prime Minister referred to the Agricultural Trust. The time has come when the Minister of Agriculture should say further that I appreciated very much the sentiments and the tone of yesterday's speeches and in particular that of the hon. and learned Member for South Antrim (Mr. Ferguson). The sentiments of the Prime Minister's speech were equally acceptable but I would hope that the new intake and the new faces on the opposite Benches will strengthen the possibility that for the business of the House to be conducted in an orderly manner?

Captain Brooke: Thank you very much, Mr. Speaker. That will make life much easier for me.

I should like to ask my right hon. and gallant Friend what measures he can take to protect our pig industry. This is something very close to my heart as we have the Ulster Swift factory in Enniskillen. Great damage is being done to us. Indeed, this smuggling is being subidised in a way by the British taxpayer. Something should be done to stop it.

The Prime Minister referred to the Agricultural Trust. The time has come when the Minister of Agriculture should throw a little more light on its actions. No doubt it has been very wise to keep quiet until it has achieved something. It announced that it was going to start a grass-drying plant and it has also done quite a lot with tax. We in this House should now have a

The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge.
<table>
<thead>
<tr>
<th>Location</th>
<th>Latitude 1</th>
<th>Longitude 1</th>
<th>Latitude 2</th>
<th>Longitude 2</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cookstown</td>
<td>54°39'N</td>
<td>6°45'W</td>
<td>54°40'N</td>
<td>6°40'W</td>
<td>3</td>
</tr>
<tr>
<td>Fermanagh</td>
<td>54°20'N</td>
<td>7°30'W</td>
<td>54°20'N</td>
<td>7°30'W</td>
<td>2</td>
</tr>
<tr>
<td>Donegal</td>
<td>54°55'N</td>
<td>8°0'W</td>
<td>54°39'N</td>
<td>8°7'W</td>
<td>3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>54°0'N</td>
<td>4°0'W</td>
<td>17°4'S</td>
<td>30°12'E</td>
<td></td>
</tr>
<tr>
<td>Antrim</td>
<td>54°42'N</td>
<td>6°12'W</td>
<td>54°40'N</td>
<td>6°15'W</td>
<td>3</td>
</tr>
<tr>
<td>Portadown</td>
<td>54°25'N</td>
<td>6°27'W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>40°0'N</td>
<td>4°0'W</td>
<td>30°52'N</td>
<td>83°32'W</td>
<td>6</td>
</tr>
<tr>
<td>County Louth</td>
<td>53°50'N</td>
<td>6°30'W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>54°0'N</td>
<td>4°0'W</td>
<td>52°10'N</td>
<td>0°42'W</td>
<td>3</td>
</tr>
<tr>
<td>United States</td>
<td>39°46'N</td>
<td>98°30'W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russia</td>
<td>60°0'N</td>
<td>100°0'E</td>
<td>41°2'N</td>
<td>74°32'W</td>
<td>4</td>
</tr>
<tr>
<td>Genoa</td>
<td>40°47'N</td>
<td>81°28'W</td>
<td>35°20'N</td>
<td>78°2'W</td>
<td>3</td>
</tr>
<tr>
<td>the Hague</td>
<td>52°5'N</td>
<td>4°18'E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scotland</td>
<td>56°0'N</td>
<td>4°0'W</td>
<td>39°2'N</td>
<td>77°10'W</td>
<td>3</td>
</tr>
<tr>
<td>Toggenburg</td>
<td>47°15'N</td>
<td>9°10'E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danegdaen</td>
<td>56°20'N</td>
<td>6°32'W</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problems:

- Identification of place names (as opposed to person names)
- Disambiguation of place names (e.g. Belfast, Antrim versus Belfast, Maine)
- Document structure - inevitably affects how the Geoparser works with individual corpora
- Lack of standardized way of dealing with georeferencing
- Only point data
HOUSE OF COMMONS.
Thursday, 18th May, 1922.

The House met at Three of the clock.

INTERIM GRANT TO SECONDARY SCHOOLS.

1.

Dr. MORRISON asked the Minister if he could not give the House before the end of the current year a consideration of the grant for secondary schools.

Dr. MORRISON moved the adjournment of the debate.

2.

Dr. MORRISON asked the Minister if he could not give the House a further consideration of the grant for secondary schools.

The MINISTER OF HOME AFFAIRS will answer the question on Monday next provided that the business of the House is not more than half finished.

1. Mr. McKean. This grant has not yet been paid and cannot be paid in the current year. The Ministry's Estimates for the current year have been considered by the House, and the six months ended 31st January last, will be paid very shortly.

2. The Resolution referred to, which, I think, calls attention to a very important point.
Defining spatial footprints
Point data is problematic...
• ‘Enforced crispness’

• The camera (or the geovisualization) never lies

• Some attempts to improve this model, e.g. anchor theory, buffering procedures
Other applications
Hephaistion

An ancient temple, attested during the Classical and Roman periods (modern location: Yanar Taş). It was known in antiquity by the name: Hephaistion.

Modern Location: Yanar Taş

Features:
- Hephaistion @ Point 11616434

Alternate representations:
- Atom
- KML
How do we get more out of digitization?

- Not just about ‘linear’ reading
- Need for authoritative cross-domain vocabularies and gazetteers, FTTs etc
- Trusted repositories
- Linking between resources
- **Useful** and **useable** interfaces