Digitizing the Canadian Parliamentary Debates

Kaspar Beelen, Christopher Cochrane, Graeme Hirst, Nona Naderi, Ludovic Rheault, Tanya Whyte

University of Toronto
Department of Computer Science
Department of Political Science
Dilipad Project Background

Digging into Linked Parliamentary Data (Dilipad) - Tri-National Project

University of Toronto
University of Amsterdam
Institute of Historical Research
Digging into Linked Parliamentary Data (**Dilipad**) - Funded by the Digging into Data Challenge
Dilipad Project Objectives

- Data Creation: Collect, Digitize and Enrich Parliamentary Proceedings in a uniform format
Dilipad Project Objectives

- Data Creation: Collect, Digitize and Enrich Parliamentary Proceedings in a uniform format

- The Netherlands: Tweede Kamer and Senaat (1815-present)
Dilipad Project Objectives

- Data Creation: Collect, Digitize and Enrich Parliamentary Proceedings in a uniform format
  - The Netherlands: Tweede Kamer and Senaat (1815-present)
  - United Kingdom: House of Commons (1803-present)
Dilipad Project Objectives

- Data Creation: Collect, Digitize and Enrich Parliamentary Proceedings in a uniform format
  - The Netherlands: *Tweede Kamer* and *Senaat* (1815-present)
  - United Kingdom: House of Commons (1803-present)
  - Canada: House of Commons (1900-present)
Dilipad Project Objectives

- Outreach: Lipad.ca
- Indexed and searchable version of the corpus (see the following presentation by Tanya Whyte)
Digitization:
Overview
FEBRUARY 25, 1902

101

The Hon. Gentleman corrobated every word I said; instead of exaggerating the statement, he made it as mild as possible. Here was a case of small-pox in a case coming to our shores where an epidemic might have been spread all over Prince Edward Island and the government of this country, or in Ireland, where the patient had come through a part of the hon. Gentleman's riding of King's, and a part of the riding of Kent County from Georgetown to a point near Charlottetown, a distance of 30 miles. The man went in that carriage or anything being provided to notify the people along the line of the arrival of the first case of small-pox that was in the country. The hon. Gentleman says that the carriage broke down and that they had to go to a blacksmith's shop to repair it. The statement is exaggerated at all, and the hon. Gentleman has himself confirmed his statement that the patient having the smallpox was driven thirty miles through the province of Prince Edward Island, liable to spread the disease all over the country, and yet the hon. Gentleman makes light of it. The public health of our people is of great importance, and every precaution should be taken to safeguard it. When the hon. Gentleman gets up here and says I have exaggerated, he says what is not correct.

Mr. HUGHES (King's). What would you do with the patient under the circumstances?

Some hon. MEMBERS. Kill him.

Mr. HUGHES (King's). Let me make a further statement. There was some local discussion about this, and the patient stationed himself in the newspaper and he was quite comfortable during the drive, that every precaution was taken, that he had no fault to find with the means by which he was brought to his hospital where he had the best medical treatment, and the proof of his having well cared for is that he recovered in a short time, and that the disease did not spread.

Mr. TAYLOR. Last year this same vote was asked for and also a supplementary vote. Will the $10,000 cover the whole service this year or will the minister ask for more?

The MINISTER OF AGRICULTURE. I hope that this will be sufficient, but that will depend entirely upon the extent of the outbreak of the disease, which I cannot foresee.

Mr. TAYLOR. In 1892 or 1893, $25,000 was asked for this same service and this was three times the amount to be voted. How is it possible that there could be such an increase in this item?

The MINISTER OF AGRICULTURE. In the first place, the work in connection with the frontier was not in existence at all in 1892, and in the second place we had not a small-pox outbreak, and we had not the known to look after. All those matters have caused the increase.

Mr. TAYLOR. I think the real explanation is that the minister had a number of doctors that he wanted to give $1,000 a month to, in order that they might draw the money and do nothing. I suppose the minister will give us the names of all the doctors that are employed in the department so that we may know where this extra $10,000 went.

The MINISTER OF AGRICULTURE. I will get that information.

Mr. MONK. Did we spend the entire $10,000 last year?

The MINISTER OF AGRICULTURE. The $10,000 is for the current year, and I am afraid we will have to ask for a supplementary vote.

Mr. MONK. Was the amount we spent last year?

The MINISTER OF AGRICULTURE. $0.

Mr. MONK. And you are asking for $5,000 more.

The MINISTER OF AGRICULTURE. Yes, and we will have to ask for more still. I regret to say that this expenditure by my department is only a small part of what the people of Canada are losing, paying on account of this small-pox outbreak. If the hon. gentleman will refer to the expenditure by the local and municipal authorities, he will find that probably several million dollars are expended in Canada this year on account of this outbreak of small-pox.

Mr. MONK. But the damage small-pox may cause has nothing to do with the inspection along the frontier. The minister tells us that there is no danger from the New England states, and must the only danger that we need think of be thousands of dollars in addition to what we spent last year in a very large amount, when the prospect is that we will not require so much. I think that vote ought to be reduced.

The MINISTER OF AGRICULTURE. The hon. gentleman does not seem to have followed the discussion. The expenditure along the frontier has been increasing but constantly increasing, and the prospect seems to be that it will go on. If it does not go on the money will not be spent, but I fear that it is necessary to provide for this contingency, and this year not only will I have to spend $10,000, but I shall have to ask for a considerable amount more to cover the expenditure for the next four months of the current year. A large amount of this money has been spent in the Yukon and in the territories dealing with small-pox.
Enrichment: Why Semantic Annotation?
Enrichment:

Why Semantic Annotation?

Mr. HUGHES (Kings). What would you do with the patient under the circumstances?

Some hon. MEMBERS. Kill him.

Mr. HUGHES (Kings'). Let me make a further statement. There was some local discussion about this, and the patient stated in the newspapers that he was quite comfortable during the drive, that every precaution was taken, that he lost no sleep and that he recov-
ered in short time, and that the disease did not spread.

Mr. TAYLOR. Last year this same vote was asked for and also a supplementary vote. The $110,000 cover the whole service this year or will the minister ask for more?

The MINISTER OF AGRICULTURE. I hope that this will be sufficient, but that will depend entirely on the extent of the outbreak of the disease, which I cannot forecast.

Mr. TAYLOR. In 1862-3, $55,000 was all that was asked for this service, but the year before that amount was to be spent. Is it possible that there could be such an increase in this item?

The MINISTER OF AGRICULTURE. In the first place, the work in connection with the frontier was not in existence at all in 1863, and in the following year, on account of these additions, we had not the funds to meet the cost of a small-pox outbreak, and we had not the funds to meet the cost of a small-pox outbreak, and we had not the funds to meet the cost of a small-pox outbreak.
Enrichment: Why Semantic Annotation?
Enrichment: Why Semantic Annotation?

Mr. HUGHES (Kings). What would you do with the patient under the circumstances?

Some hon. MEMBERS. Kill him.

Mr. HUGHES (King’s). Let me make a further statement. There was some local discussion about this, and the patient stated himself in the newspapers that he was quite comfortable during the drive, that every precaution was taken, that he had no fault to find with the means to bring him to the hospital where he would get the best medical treatment, and the proof of his being well cared for is that he recovered in a short time, and that the disease did not spread.

Mr. TAYLOR. Last year this same vote was asked for and also a supplementary vote. Will the $110,000 cover the whole service this year or will the minister ask for more?

The MINISTER OF AGRICULTURE. I hope that this will be sufficient, but that will depend entirely on the extent of the outbreak of the disease, which I cannot foretell.
Enrichment: Why Semantic Annotation?

Mr. HUGHES (Kings). What would you do with the patient under the circumstances?

Some hon. MEMBERS. Kill him.

Mr. HUGHES (King’s). Let me make a further statement. There was some local discussion about this, and the patient stated himself in the newspapers that he was quite comfortable during the drive, that every precaution was taken, that he had no fault to find with the means to bring him to the hospital where he would get the best medical treatment, and the proof of his being well cared for is that he recovered in a short time, and that the disease did not spread.

Mr. TAYLOR. Last year this same vote was asked for and also a supplementary vote. Will the $110,000 cover the whole service this year or will the minister ask for more?

The MINISTER OF AGRICULTURE. I hope that this will be sufficient, but that will depend entirely on the extent of the outbreak of the disease, which I cannot foretell.
Corpus Structure

Diagram showing the structure of a corpus, with nodes labeled as "Procedings Corpus", "Proceedings File", "Speech", "Parties Corpus", and "Members Corpus". Arrows indicate relationships between these nodes.
Corpus Structure

Proceedings
Dilipad Scheme

Proceedings
Dilipad Scheme

Proceedings
Dilipad Scheme

Proceedings
Dilipad Scheme

Proceedings
Dilipad Scheme

Proceedings
Dilipad Scheme

Proceedings
Dilipad Scheme

Members
Dilipad Scheme

Members
Dilipad Scheme

Members
Dilipad Scheme

Parties
Project Workflow

Scans → OCR conversion → Plain Text → Data Enrichment → Merge

- Link Entities
- Detect Entities

DBpedia
Authority Lists

Collect External Information

Add to Proceedings to Linked Open Data Cloud

Annotated XML
Project Workflow

1. OCR Conversion

Scans → OCR conversion → Plain Text

Data Enrichment

Link Entities → Merge

Detect Entities

Collect External Information

Annotated XML

Add to Proceedings to Linked Open Data Cloud

DBpedia

Authority Lists
Project Workflow

1. Scans → OCR conversion → Plain Text

2. Link Entities → Data Enrichment → Detect Entities

3. Collect External Information → Merge

4. DBpedia

5. Authority Lists

6. Annotated XML

7. Add to Proceedings to Linked Open Data Cloud

2. Structuring Text
Project Workflow

3. Linking Data

31
Step 1: OCR Conversion

1. Scans -> OCR conversion -> Plain Text
2. Data Enrichment
   - Link Entities
     - DBpedia
     - Authority Lists
   - Collect External Information
3. Detect Entities
4. Merge
5. Add to Proceedings to Linked Open Data Cloud

Annotated XML
DEFENCE EXPENDITURE
APPOINTMENT OF SPECIAL COMMITTEE
The House Resumed [...] 
Mr. G. C. Nowlan (Annapolis-Kings):  
Mr. Speaker, I intend to intervene but briefly in his debate. [...] 
Mr. Jean Francois Pouliot (Temiscouata):  
Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. [...]
Step 2: Structuring Text

1. Scans → OCR conversion → Plain Text
2. Data Enrichment
3. Link Entities
4. Collect External Information
5. Merge
6. Add to Proceedings to Linked Open Data Cloud
7. Detect Entities
8. Annotated XML

Additional resources:
- DBpedia
- Authority Lists
Mr. G. C. Nowlan (Annapolis-Kings): Mr. Speaker, I intend to intervene but briefly in his debate. [...] 
Mr. Jean Francois Pouliot (Temiscouata): Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. [...]
DEFENCE EXPENDITURE
APPOINTMENT OF SPECIAL COMMITTEE

The House Resumed

Mr. G. C. Nowlan (Annapolis-Kings):
Mr. Speaker, I intend to intervene but briefly in his debate. […]

Mr. Jean Francois Pouliot (Temiscouata):
Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. […]

Matching Patterns with **Regular Expressions**

E.g. Wildcards, Canad* = Canada, Canadian,
DEFENCE EXPENDITURE
APPOINTMENT OF SPECIAL COMMITTEE

The House Resumed

Mr. G. C. Nowlan (Annapolis-Kings):
Mr. Speaker, I intend to intervene but briefly in his debate. [...]

Mr. Jean Francois Pouliot (Temiscouata):
Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. [...]

Identifying Patterns

\(\text{\texttt{(\n[A-Z]\{3,\}\n)}\)
DEFENCE EXPENDITURE
APPOINTMENT OF SPECIAL COMMITTEE

The House Resumed

Mr. G. C. Nowlan (Annapolis-Kings):
Mr. Speaker, I intend to intervene but briefly in his debate. [...] 

Mr. Jean Francois Pouliot (Temiscouata):
Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. [...] 

(\nMr\.|s[A-Za-z]+s(.?)):
Issues: Variation

Hon. Jean J. Charest (Minister of State (Youth) and Minister of State (Fitness and Amateur Sport and Deputy Leader of the government in the House of Commons)):

Mr. Rompkey:
Issues: Variation

^((?:Sir|M\.|Mr\.|Mr\,|Hon\.|The\sHon\.|Right\sHon\.|The\sRight\sHon\.|Miss|Mrs\.|Ms\.)\s(?:[A-Zdv][\-\w\.']{1,25}\s{0,1})\s{0,1}(?:\(.+?)\s{0,1}(?:\:|;))\s{0,1}(?:\:(?:\:moved:)\s|\s\:(?:moved):|:;))

Etcetera...
Issues: Changes over time

Mr. JAMIESON. If the hon. gentleman will excuse me,—

Mr. MILLS (Bothwell). I do not intend that the hon. gentleman shall interrupt me. If he wishes to make any more incorrect statements, he can take some other occasion to make them. There was a local election in the hon. gentleman’s constituency, and the president of the prohibition committee, and a member of the Temperance Alliance, was a candidate.

1888
Issues: Changes over time

yesterday and I did not see the report of what Premier Manning said—

**Mr. Diefenbaker:** Did it happen?

**Mr. St. Laurent:** —but I told him that I would be very glad to sit in with the representatives that he mentioned as being those he was going to invite. I told him I was sure

1955
Issues: OCR Errors

BAILLANiTYNE
BALLAINTYNE
BALLANT1NE
BALLAiNTYNE
iBALiLANTYNE
Issues: OCR Errors

BALLANTYNE
BALLANTYNE
BALLANTINE
BALLANTYNE

\[ \text{BALLANTYNE} = \text{BALLANTYNE} \]
Mr. G. C. Nowlan (Annapolis-Kings):
Mr. Speaker, I intend to intervene but briefly in his debate. […]

Mr. Jean Francois Pouliot (Temiscouata):
Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. […]
Annotating the Proceedings

```xml
<preproceedings>
<topic_start>DEFENCE EXPENDITURE</topic_start>  
<topic_start>APPOINTMENT OF SPECIAL COMMITTEE</topic_start>  
<stage-direction>The House Resumed [...]</stage-direction>  
<speech_start>Mr. G. C. Nowlan (Annapolis-Kings):</speech_start>  
<p>Mr. Speaker, I intend to intervene but briefly in his debate. [...]</p>  
<speech_start>Mr. Jean Francois Pouliot (Temiscouata):</speech_start>  
<p>Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. [...]</p> 
</proceedings>
```
Annotating the Proceedings

<pm:proceedings pm:id="ca.proc.d.1953-01-20">
  <pm:topic pm:id="ca.proc.d.1953-01-20.1" pm:title="DEFENCE EXPENDITURE">
    <pm:scene pm:id="ca.proc.d.1953-01-20.1" pm:title="APPOINTMENT OF SPECIAL COMMITTEE" pm:type="topic">
      <pm:speech pm:id="ca.proc.d.1953-01-20.1.1" pm:speaker="Mr. G. C. Nowlan (Annapolis-Kings)">
        Mr. Speaker, I intend to intervene but briefly in his debate.
      </pm:speech>
    </pm:scene>
  </pm:topic>
  <pm:topic pm:id="ca.proc.d.1953-01-20.2" pm:title="DEFENCE EXPENDITURE">
    <pm:speech pm:id="ca.proc.d.1953-01-20.2.1" pm:speaker="Mr. Jean Francois Pouliot (Temiscouata)">
      I do not intend to speak today only as a member of parliament or as a member of the Liberal party.
    </pm:speech>
  </pm:topic>
</pm:proceedings>
Mr. Speaker, I intend to intervene but briefly in his debate.

I do not intend to speak today only as a member of parliament or as a member of the Liberal party.
Step 3: Linking Data

Diagram:
- Scans → OCR conversion → Plain Text
- Data Enrichment
  - Detect Entities
  - Annotated XML
- Link Entities
  - DBpedia
  - Authority Lists
- Collect External Information
- Merge
- Add to Proceedings to Linked Open Data Cloud
Disambiguating Entities

Mr. Jean Francois Pouliot (Temiscouata)

I do not intend to speak today only as a member of parliament or as a member of the Liberal party.
Disambiguating Entities

Mr. Jean Francois Pouliot (Temiscouata)

I do not intend to speak today only as a member of parliament or as a member of the Liberal party.
Disambiguating Entities

Mr. Jean Francois Pouliot (Temiscouata)
Adding Information

Mr. Jean Francois Pouliot (Temiscouata)

I do not intend to speak today only as a member of parliament or as a member of the Liberal party.
Adding Information

Mr. Jean Francois Pouliot (Temiscouata)
Project Summary

Project Output:
- Structured and enriched parliamentary corpus. Includes all House of Commons debates from 1900 to present.
- Linked to ParlInfo and other knowledge sources such as Wikipedia.
- Easy-to-use and flexible search engine (Lipad.ca).
Dilipad Team

Canada

- Team: Kaspar Beelen, Chris Cochrane, Graeme Hirst, Nona Naderi, Ludovic Rheault, Tanya Whyte
- Interns: Tim Alberdingk-Thijm, Mike Kimmins, Roman Polyanovsky

Netherlands

- Team: Jaap Kamps, Maarten Marx
- Other Contributors: Hosein Azarbonyad, Mostafa Denghani, Alex Olieman
- Interns: Kees Halvemaan, Sander Lijbrink

United Kingdom

Funding Agencies

- Social Sciences and Humanities Research Council (CAN)
- National Sciences and Engineering Research Council (CAN)
- Canada Foundation for Innovation (CAN)
- Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NED)
- Arts and Humanities Research Council (UK)
- Economic and Social Research Council (UK)
- National Endowment for the Humanities (USA)
- National Science Foundation (USA)
- Institute of Museum and Library Services (USA)
- Joint Information Systems Committee (UK)