A metadata infrastructure for the analysis of parliamentary proceedings

Richard Gartner
Centre for e-Research, Department of Digital Humanities
King's College London
London, United Kingdom
richard.gartner@kcl.ac.uk

***Abstract*—This work-in-progress article discusses DILIPAD (Digging into Linked Parliamentary Data), a project funded under the Digging Into Data Challenge. DILIPAD aims to create an extensive corpus of structured XML data of parliamentary proceedings from three countries (United Kingdom, Netherlands and Canada) in order to enable large-scale diachronic analyses of their content. The corpora integrate the textual data of proceedings within contextual metadata encoded in the XML schema Parliamentary Metadata Language (PML). The article discusses the background to the project, the construction of the corpora and highlights they ways in which they may be used for quantitative and qualitative analysis.**

Keywords—metadata; corpus analysis; parliamentary history; XML

I. Introduction

Although much of the data which first brought the concept of Big Data to attention originated in the sciences, its applicability to the humanities is currently being explored in greater depth than previously. The concept itself is subject to a variety of definitions, but most tend to agree on the distinctive features highlighted by Ward and Baker [1, p. 2]: size, complexity and technology (the last being the development and use of tools capable of processing large, complex datasets). Although the datasets in humanities Big Data projects are often smaller than those originating in the sciences [2, p. 462], its complexity and the need for new tools and techniques to handle this are just as demanding.

Meeting these challenges in the main rationale behind the "Digging into Data Challenge"[3].This "challenge", an open competition for innovative projects in large-scale data analysis in the humanities and social sciences, has funded the DILIPAD (Digging into Linked Parliamentary Data) [4] project which is the subject of this work-in-progress paper. This project aims to develop new methodologies for the qualitative analysis of large volumes of records of legislative proceedings in three countries (United Kingdom, Canada and the Netherlands). It is attempting to do so by producing structured text corpora from these records and devising techniques, analogous to those already used in corpus linguistics, to analyse these across temporally and geographically diverse ranges of data.

II. Background

A. Digitizing Parliamentary Data

A large corpus of parliamentary proceedings has been available in all three countries covered by the Dilipad project for some years. In the United Kingdom, the full text of proceedings has been digitized from 1803 onwards in a number of notable projects, including British History Online [5] and the UK Parliament's own conversion of the Hansard record of parliamentary debates [6]. In the Netherlands, two hundred years of proceedings have been digitized [7, p. 3671] as part of *DutchParl,* a corpus of Dutch-language proceedings [7]. In Canada, scanned images of proceedings date as far back as 1867 [7] although machine-readable texts of these only go back to the 1990s.

These diverse projects have produced large collections of machine-readable texts, but these conform to a diverse set of encoding standards which renders large-scale analysis of their contents difficult to achieve. The UK's Hansard project, for instance, uses XML files conforming to in-house schemata which undergo several revisions over the period of coverage. The Netherlands data conform to an in-project schema, Politicalmashup [9], while the Canadian data conforms to TxtMap, a schema devised in-house to represent the text of single page images with limited semantic content [9]. Other projects, such as a recent project to scan six years of proceedings from the Estonian parliament, use the TEI (Text Encoding Initiative) [10].

This heterogeneity of approaches to encoding limits the analytical potential of each collection and severely curtails their potential for cross-collection analysis. Generic bibliographic metadata schemas, such as the TEI Header or the more sophisticated MODS (Metadata Object Description Schema), do not offer sufficiently specific semantics to describe these proceedings adequately. Of the bespoke XML applications used by the projects noted above, only the Politicalmashup schema is devised specifically for textual analysis, and this is itself limited to a narrow range (specifically the structure of proceedings in terms of speeches, interruptions, interventions etc). To allow more sophisticated querying of the record and qualitative analyses to be carried out requires the use of a new more generic schema.

B. LIPARM (Linking the Parliamentary Record Through Metadata)

A recent attempt to address these issues is the LIPARM project, based at King's College London, which aimed to establish a common metadata format for parliamentary (and more generally legislative) proceedings [11]. The project, which finished in 2013, produced an XML schema, named PML (Parliamentary Metadata Language) which defined core facets of the record and a network of semantic links between them [12]. Seven such high-level facets are defined in PML as follows:-

Unit name	Example from Canadian Parliament
Units	Parliament House of Commons Senate
Functions	Prime Minister Speaker of the Senate
Persons	Stephen Harper
Calendar objects	Parliament 2008-2011
Proceedings groups	Canadian Environmental Assessment Act, 2003
Proceedings objects	Forestry Industry Support Debate, 31 March 2008
Vote events	Division taken on 15 January 1980

TABLE I. PML FACETS

These broad facets are qualified, so refining their semantic coverage, by the use of persistent *type* and *typeURI* attributes:

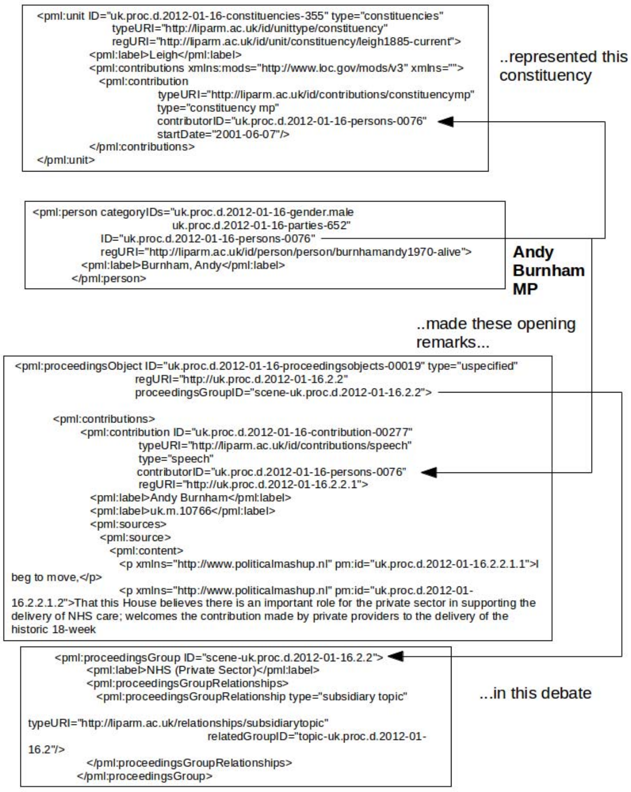
```
<unit type="constituency"
      typeURI="http://liparm.ac.uk/id/unittype/consituency">
```

where *typeURI* references an entry in a controlled vocabulary or ontology.

Semantic links are made between components by the labelling of every component with an XML ID and the pervasive use of attributes of type IDREF to reference these. A network of links around a single Member of Parliament (MP), for instance, may take the form shown in Fig. 1. Here an MP is linked to the constituency he serves, the remarks he made in a session and the debate within which these remarks were made.

The LIPARM project concerned itself primarily with creating a new resource-discovery tool for the complex contents of parliamentary proceedings rather than an analytical tool for this dataset. As part of the project, for instance, a simple prototype interface was created [12, p. 33] which provides faceted browsing of the components encoded within PML files and links to the full text of the proceedings themselves.

Nonetheless, it was recognised that discovery *per se*, even in the form that PML allowed (which was notably more sophisticated than had previously been possible), did not realise the full potential of this new schema. The network of semantic linkages encoded within its architecture could readily form the basis of a powerful analytical tool for this body of data. The structured semantic framework provided by PML allows the possibility of creating datasets or corpora of parliamentary data amenable to the analytical techniques which are becoming increasingly available as 'big-data' methodologies develop.



```
<pml:unit ID="uk.proc.d.2012-01-16-constituencies-355" type="constituencies"
          typeURI="http://liparm.ac.uk/id/unittype/constituency"
          regURI="http://liparm.ac.uk/id/unit/constituency/leigh1885-current">
      <pml:label>Leigh</pml:label>
      <pml:contributions xmlns:mods="http://www.loc.gov/mods/v3" xmlns="">
        <pml:contribution
                  typeURI="http://liparm.ac.uk/id/contributions/constituencymp"
                  type="constituency mp"
                  contributorID="uk.proc.d.2012-01-16-persons-0076"
                  startDate="2001-06-07"/>
      </pml:contributions>
</pml:unit>
```

..represented this constituency

```
<pml:person categoryIDs="uk.proc.d.2012-01-16-gender.male
                         uk.proc.d.2012-01-16-parties-652"
        ID="uk.proc.d.2012-01-16-persons-0076"
        regURI="http://liparm.ac.uk/id/person/person/burnhamandy1970-alive">
      <pml:label>Burnham, Andy</pml:label>
    </pml:person>
```

Andy Burnham MP

..made these opening remarks...

```
<pml:proceedingsObject ID="uk.proc.d.2012-01-16-proceedingsobjects-00019" type="uspecified"
                  regURI="http://uk.proc.d.2012-01-16.2.2"
                  proceedingsGroupID="scene-uk.proc.d.2012-01-16.2.2">

      <pml:contributions>
           <pml:contribution ID="uk.proc.d.2012-01-16-contribution-00277"
                     typeURI="http://liparm.ac.uk/id/contributions/speech"
                     type="speech"
                     contributorID="uk.proc.d.2012-01-16-persons-0076"
                     regURI="http://uk.proc.d.2012-01-16.2.2.1">
              <pml:label>Andy Burnham</pml:label>
              <pml:label>uk.m.10766</pml:label>
              <pml:sources>
                <pml:source>
                  <pml:content>
                    <p xmlns="http://www.politicalmashup.nl" pm:id="uk.proc.d.2012-01-16.2.2.1.1">I
beg to move,</p>
                    <p xmlns="http://www.politicalmashup.nl" pm:id="uk.proc.d.2012-01-
16.2.2.1.2">That this House believes there is an important role for the private sector in supporting the
delivery of NHS care; welcomes the contribution made by private providers to the delivery of the
historic 18-week
```

```
      <pml:proceedingsGroup ID="scene-uk.proc.d.2012-01-16.2.2">
            <pml:label>NHS (Private Sector)</pml:label>
            <pml:proceedingsGroupRelationships>
              <pml:proceedingsGroupRelationship type="subsidiary topic"

typeURI="http://liparm.ac.uk/relationships/subsidiarytopic"
                                        relatedGroupID="topic-uk.proc.d.2012-01-
16.2"/>
            </pml:proceedingsGroupRelationships>
          </pml:proceedingsGroup>
```

...in this debate

Fig. 1. Sample linkages encoded in PML

III. DILIPAD (DIGGING INTO LINKED PARLIAMENTARY DATA

Realising the analytical potential of PML is one of the prime rationales behind the DILIPAD project introduced here. This project, conceived from its inception as a follow-one to LIPARM, concentrates specifically on the creation of large corpora of PML-encoded texts and the development of analytical techniques for interrogating them on large scales. As part of this, the project is defining core historical research questions which are amenable to this type of analysis and the methodologies for extraction and analysis of corpus data to answer these.

A. Corpus creation

The initial stages of the project, still current at the time of writing, are concerned specifically with the creation of the corpus of structured data in PML format. The source materials used, although all in XML, are diverse in their content, the architectures of the schemata to which they conform, the identifier schemes employed (for instance for persons) to enable linkages between components and their granularity. For these reasons, each type of source material requires a specific mapping to the architecture of PML.

For the United Kingdom material, it was decided to convert data not from the relatively inconsistent Hansard dataset offered by the UK Parliament, but from an independent service *TheyWorkForYou.com* [13]. This service, designed to offer UK citizens access to information on the contributions made by their MPs to parliamentary business, has produced structured XML extracted from Hansard proceedings as far back as 1935. These data are available in the logical and consistent *politicalmashup* schema within which Dutch data from the *DutchParl* corpus are also encoded; data for both of these parliaments will therefore be extracted from files encoded in this schema. The Canadian data are not taken directly from *TxtMap*-encoded files, but from a more structured body of XML data extracted from them which conforms to a simple schema devised by the Canadian Hansard publication unit.

The conversion itself is achieved using *XSLT (eXtensible Stylesheet Transformation)* transforms. This is a relatively complex, three-stage process as follows:-

- the initial stage extracts all relevant data from the original dataset using the date of proceedings as its primary limiting criterion: it also extracts relevant data, based on this date, from auxiliary controlled vocabularies (for instance, lists of members or constituencies)
- the second stage populates the component elements of the newly created XML instances with internal IDs to act as referents for semantic links
- the third stage creates the semantic links themselves

These XSLT transformations have now been tested on the original corpora and proven effective in creating the accurate semantic links which are essential for this methodology to produce viable analytical results. Crucial to this is the application of consistent and logical identifiers for key components. Where these are present in the source materials this is readily achieved: where these are absent, the conversion is less certain.

An example of the problems arising from a lack of identifiers of this type is the encoding of votes. PML has an extensive element set for the recording of votes taken as part of proceedings, including a facility for listing those members who vote for each option: for example:-

```
<option regURI="http://liparm.ac.uk/id/votingoption/no">
    <vote>Adams, Mrs Irene</vote>
    <vote voterID="debates1994-11-02a-persons-0010">Ainger, Nick</vote>
    <vote>Callaghan, Jim</vote>
</option>
```

The <vote> element shown here allows for the inclusion of a *voterID* attribute which may be used to reference the <person> element within the PML file for the member who is casting their vote. As can be seen, however, this can only be generated for a limited number of votes cast owing to the lack of unambiguous IDs in the original source:-

```
<tr>
    <td align="center" colspan="2"><b>NOES</b></td>

</tr>
<tr>
    <td>Adams, Mrs Irene</td>
</tr>
<tr>
    <td>Ainger, Nick</td>
    <td>Callaghan, Jim</td>
</tr>
```

The vote is here recorded as a simple HTML-style table with no identifiers for each member. The assignment of PML identifiers therefore can be done by nothing more sophisticated than simple string matching, which results in a limited number of successful ID assignments (approximately 60% for more recent materials, 30% for their older counterparts).

Much of the work currently underway in the project is concerned with resolving issues of this type: extensive data cleansing is needed, some of which can be automated but much of which requires labour-intensive manual intervention. This is required particularly to establish the unambiguous identification for individuals and other key components of the record. Once as much of this work as is feasible within the project timescale is completed, complete corpora from 1935 will be generated to form the base dataset for the remaining stages of the project.

B. Corpus analysis

The tightly-structured PML architecture and its network of semantic links (both within the PML file itself and beyond it) allow the textual content of the parliamentary record and its associated metadata to be co-analysed using standard XML search and processing methods. The next stage of the project after the corpus creation will be the design of a set of query statements for interrogating it analytically.

At this early stage, a preliminary set of potential queries have been drawn up by a number of parliamentary historians. These include such questions as:-

- How many times did XMP speak in session 1939-40 or in calendar year 1939?
- How many times did XMP intervene in someone else's speech in the same period?
- How many times did XMP ask a Question (in Question time) in Y period?
- How many times did XMP ask a Question about Z subject, or direct a question to Department of AB, or to the Secretary of State for AC?
- How many times did XMP vote?
- Give me a list of what XMP spoke about in Y period
- Give me a list of what XMP voted on in Y period (and ideally how he voted, aye or no)
- Give me a list of all debates/votes/bills relating to agriculture (or subject K) in Y period [in which XMP spoke]
- Give me a list of all bills debated in Y period
- Give me a list of all MPs who used the word 'asylum' in debates in Y period

All of these are readily resolved by interrogating a corpus of PML-encoded data using XQUERY queries. For example a list of MPs who used the word 'asylum' in debates in a given period is given by:-

```
{for $match in
    //pml:contribution[@type="speech"]//pm:p/text()[contains(.,'asylum')]
    return
    {$match/ancestor::pml:contribution/@contributorID}

}
```

More complex, and potentially more interesting, analyses become possible when these relatively simple XQUERY extraction techniques are conjoined with more sophisticated methodologies derived from pre-existing practices in text-mining and corpus linguistics. In particular the well-established fields of discourse and sentiment analysis can be applied to the methodologies devised here to examine, for instance, the role of gender in the parliamentary discourse (as has recently been attempted on a smaller scale for the Swedish parliament [14]).

The project will only begin to examine these questions and techniques in depth in 2015 after the construction of the corpora and the methodologies for data extraction and recombination are fully defined. Nonetheless, the ways in which quantitative text-mining techniques may be allied to more qualitative analytical methods of this type are already emerging.

The primary methodology that will be employed will be to link standard corpus-based linguistic analyses (such as *n-gram*-generated results) with the contextual semantic linkages encoded in PML. In this way, the standard data that emerge from such analyses (such as basic frequency counts, but also including more sophisticated material such as collocation data) will be linked with the semantic metadata linkages encoded within the PML architecture. Some possible lines of enquiry amenable to analysis of this kind might involve the evolution of language related to immigration (for instance, the rise and fall of racialist language) in a long-term diachronic perspective and across the parliaments of the three countries covered by the project.

The particular strength of the DILIPAD approach to such study is the ability to embed the results of these linguistic analyses in the context in which they are found. Providing a large contextual architecture within which the results of these analyses can be evaluated should allow more empirically-based conclusions to be drawn from them than would be possible from lexical analysis alone. It is in the conjoining of the lexical and the contextual that this technique offers perhaps its greatest potential contribution to historical research.

IV. Conclusions

Although this project is only one quarter into its span, it has already demonstrated that the PML schema provides a robust framework for the construction of large-scale corpora for parliamentary and legislative proceedings. These corpora, which integrate textual data with contextual metadata, can form the basis of sophisticated analyses in a manner analogous to those which are well established in such fields as corpus linguistics.

The use of PML's architectures allows, for the first-time, large-scale diachronic analyses which cross international boundaries. Such analyses are currently little used in the field of parliamentary history, owing to a paucity of semantically-interoperable data and metadata. Even at this relatively early stage, therefore, it seems likely that the this project will alter the landscape of parliamentary historiography to a substantial degree.

Acknowledgment

Research for this paper was undertaken as part of the Digging into Data Challenge, Round Three, funded in the UK by the Arts and Humanities Research Council, the Economic and Social Research Council, and Jisc.

References

[1] J. S. Ward and A. Barker, ‘Undefined By Data: A Survey of Big Data Definitions’. 2013.

[2] L. Manovich, ‘Trending: the Promises and the Challenges of Big Social Data’, in *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press, 2012, pp. 460–475.

[3] ‘Digging Into Data > Home’, 2014. [Online]. Available: http://www.diggingintodata.org/. [Accessed: 23-Jun-2014].

[4] ‘Digging into Linked Parliamentary Data’, 2014. [Online]. Available: http://dilipad.history.ac.uk/. [Accessed: 23-Jun-2014].

[5] History of Parliament Trust, ‘British History Online’, 2011. [Online]. Available: http://www.british-history.ac.uk/. [Accessed: 20-Jan-2012].

[6] M. Marx and A. Schuth, ‘DutchParl: The Parliamentary Documents in Dutch’, presented at the Seventh International Conference on Language Resources and Evaluation, Valetta, 2010, pp. 3670–3677.

[7] ‘House of Commons Debates, 1st Parliament, 1st S... - Historical Debates of the Parliament of Canada’, 2014. [Online]. Available: http://parl.canadiana.ca/view/oop.debates_HOC0101. [Accessed: 23-Jun-2014].

[8] ‘PoliticalMashup’. [Online]. Available: http://politicalmashup.nl/. [Accessed: 23-Jun-2014].

[9] Canadian.org, ‘TxtMap’. [Online]. Available: http://www.canadiana.ca/schema/2012/xsd/txtmap/txtmap.xsd. [Accessed: 23-Jun-2014].

[10] University of Tartu, ‘Reference corpus of Estonian: Transcripts of Riigikogu (Estonian Parliament)’, 2013. [Online]. Available: http://www.cl.ut.ee/korpused/segakorpus/riigikogu/index.php?lang=en . [Accessed: 02-Oct-2014].

[11] King’s College London, ‘Linking Parliamentary Records through Metadata’, 2012. [Online]. Available: http://liparm.cerch.kcl.ac.uk/. [Accessed: 23-Jan-2012].

[12] R. Gartner, ‘Parliamentary Metadata Language: an XML approach to integrated metadata for legislative proceedings’, *Journal of Library Metadata*, vol. 13, no. 1, pp. 17–35, 2013.

[13] ‘TheyWorkForYou: Hansard and Official Reports for the UK Parliament, Scottish Parliament, and Northern Ireland Assembly - done right’. [Online]. Available: http://www.theyworkforyou.com/. [Accessed: 24-Jun-2014].

[14] H. Bäck, M. Debus, and J. Müller, ‘Who Takes the Parliamentary Floor? The Role of Gender in Speech-making in the Swedish Riksdag’, *Political Research Quarterly*.