Parliamentary Metadata Meeting: a brief report

Robertson Room, Portcullis House, Westminster
June 8th 2011

Introduction

A one-day meeting on parliamentary metadata was held with financial support from JISC on June 8th 2011. Present at the meeting were a broad representation of UK projects currently working on parliamentary materials: the attendees were Richard Gartner (KCL - Chair), Lorna Hughes (National Library of Wales), Rob Newman (ProQuest), Rob Phillips (National Library of Wales), George Woodman (Northern Ireland Assembly Library), Claire Moore (ProQuest), Jessica Mulley (House of Commons), Paul Ell (Queens University Belfast), Paul Seaward (History of Parliament Trust), Simon Burrowes (Northern Ireland Assembly Official Report), Lynn Gardner (House of Commons Journal), Michael Fincham (ProQuest), Jane Winters (Institute of Historical Research), Jennie Lynch (Parliamentary Archives, Adrian Brown (Parliamentary Archives), and Liam Laurence-Smyth (Clerk of Journals, Westminster).

This meeting was called as it is widely acknowledged in the community of practitioners in the area of parliamentary papers that an approach to linking the substantial number of projects in existence would be highly valuable. This has been discussed in the past, particularly in a series of meetings convened by Paul Seaward in the mid-2000s, but this has remained something of an aspiration until the present day. This meeting aimed to revive discussion on this, and to examine a possible approach to parliamentary metadata which would make this goal feasible.

Current state of metadata practice

The projects that attended the meeting revealed a disparate set of approaches to metadata. Amongst those present, a number submitted descriptions of their current practices:-

National Library of Wales

Metadata is created for archived websites at target level (individual website) in the OPAC (using MARC and LCSH as a controlled vocabulary). Metadata is also created in the Web Curator Tool (WCT) which is used to harvest websites. The description tab allows recording of fields such as: Title, Publisher, Format, and Language (if the website is in English, Welsh, fully or partially bilingual).

ProQuest

Metadata for the House of Commons Parliamentary Papers varies by period. The 18th-century data came from the BOPCRIS digitization. The 19th century is based on the Cockton subject catalogue, which is a comprehensive and controlled listing of all papers in that period. The 20th century is based on the HMSO decennial indexes up to 1979 and thereafter on the records from POLIS and PIMS. ProQuest use the XML files created by Robert Brook for Hansard, and are also using the Rush database of MPs provided by History of Parliament.

Stormont papers

A combined index has been created based on the annual volume index compiled by Hansard officials. This has been enhanced and standardised to adapt to changes in indexing methodologies used by Hansard over the 50-year period. This provides the basis for a controlled vocabulary search functionality that can be qualified by volume(s), year(s), date(s), and index entry type. To aid searching further, Members of Parliament, offices of state, etc. have been XML tagged within the text and work is currently taking place to tag place-names.

Proposed strategy for parliamentary metadata

Richard Gartner proposed a strategy for metadata integration based on a series of discussions with interested parties.

At the core of the proposed plan is the development of a generic XML scheme for parliamentary metadata. This would provide a structure within which key elements could be identified and tracked across collections, e.g., people, items of business, bills, acts, etc. Ideally, the XML scheme should be compatible with linked data to generate RDF triples. The core idea is that all projects working with similar materials – both contemporary and historic – could point to this scheme, which would provide a centralized way of describing their component metadata. This scheme could be known as PML: Parliamentary Markup Language.

Such a scheme would rely on a logical system of identifiers both internal (to the XML file) and external (URIs): all linking would be done by these identifiers.

A key part of the strategy would be the production of a comprehensive set of controlled vocabularies which will include URIs to identify both people and things (such as bills, acts, items of parliamentary business etc). This must be machine readable, following established standards for encoding. This could possibly be made available in MADS (Metadata Authority Description Schema), a new XML schema from the Library of Congress. This is also available as an OWL ontology – OWL can be generated directly from MADS and so the authority lists can be published directly as machine-readable ontologies.

The preferred option would be to embed PML as an extension to MODS, which would then be packaged in METS; it could also function as an addition to the TEI header, as a standalone metadata record, or as linked data in semantic web applications.

Components of PML

After some discussion, a number of key components of the parliamentary metadata scheme were identified, as follows:-

- **persons:** a controlled method for referencing people, both parliamentary members and non-members, identified by persistent URIs
- **roles:** some controlled method of encoding the role of individuals is required: one possibility is a user-defined members' taxonomy with a controlled vocabulary of roles to allow cross-document searching as shown in example 1 overleaf. Members can then be linked to this taxonomy via XML IDREFs to record their roles.
- **chronologies:** a hierarchical method for identifying (in order) parliaments/sessions/sittings
- **proceedings:** a division between legislative and non-legislative proceedings is suggested.

Legislation may be described as in example 2 overleaf, including such facets as the type of legislative object and the stage in parliamentary proceedings it has reached. Non-legislation would include such items as business of the house, prayers, questions, expulsions, judicial business etc.: a taxonomy for such items already exists which could form the basis of this part of the schema.

- **divisions:** these should include details of the votes cast, which could be neatly expressed using the IDs of members, as shown in example 3

```
<members>
     <membersTaxonomy>
            <primaryCategory ID="memCat-003" type="houseMember">
                 <categoryName>Members of House</categoryName>
            </primaryCategory>
     </membersTaxonomy>
     <membersList>
            <member ID="memb-0001" categoryID="memCat-003">
                  <person reg="Anderson, Rt. Hon. Sir Robert Newton">
                         <nameFreeText>Anderson Sir R. N.</nameFreeText>
                  </person>
             </member>
      </membersList>
</members>
```

Example 1: Members' taxonomy and sample member's entry

```
<legislation>
     <legislativeObject type="bill | act | other" URI="">
        <name/>
    <stage type="first | second | committee | third | royalAssent" URI=""  ID=""/>
</legislation>
```

Example 2: Possible legislative object scheme

```
<division objectID="bill2-reading2">
          <ayes membersID="member1 member20">
          <noes members ID="">
</division>
```

Example 3: Possible encoding of divisions

Clearly this short list of potential components can only represent a small part of the total required for a working system, and so these recommendations can only form the beginnings of further discussions into this part of an overall metadata strategy.

Controlled vocabularies

A key component of a parliamentary metadata strategy would be a list of controlled vocabularies, the components of which would be identified with persistent URIs. These URIs would then be referenced from within records to allow cross-searching and browsing across systems.

A number of metadata components requiring control in this way were identified; in addition, possible sources for each vocabulary type were proposed.

- **persons**: a number of important sources already exist which could form the basis of authority lists for people – these include:-
    - The **Rush Database** used by the History of Parliament Trust and ProQuest: a few issues arise with the data itself, but this is an important resource which includes names, functions, constituencies and family, all of which have persistent IDs. This does not cover the House of Lords.
    - The **Houses of Parliament** Database covers 1386-1868 – some databases cover pre-1386 election returns.
    - **Northern Ireland:** sources include the Northern Ireland Political Directory, developed by Elliott/Flackes. Authoritative lists also appear in every Hansard, and deaths are listed.

    IPR issues on all of these sources should present few problems: many are covered by Crown copyright and permission to use others (for example the Rush Database) should be readily obtainable.
- **Judicial business: w**here the Lords functioned as a court, proceedings will be covered by law reports. Source of Taxonomy: Lords Journal index
- **Geography:** The e-Government interoperability framework (eGIF), based at the National Archives have done some work on this. In addition, the Rush Database has constituency information, and Humphrey Southall has done a constituency database as part of the Vision of Britain project. CLAIM at Ulster has done work on Ulster Place names.
- **Subjects:** Hansard indexes include subject terms which are in machine-readable form; these are not controlled to any great extent and work would be required on rationalising and mapping them
- **Offices:** The Rush database has published “Office Holders in Early Modern England”: a URI is needed for each office. Lists exist for office holders for Stormont and Wales.
- **Dates:** beyond using the standard ISO format for dates, an ontology is needed to map regnal years to parliaments to sessions to sittings: the “Handbook of British Chronology” covers this.

Conclusions

All present agreed that the approach taken here is solid and that it should be taken forward into future work. This should include the following:-

- defining and publishing the PML schema itself
- drawing up all necessary controlled vocabularies and publishing them in a variety of formats, including MADS and OWL
- establishing methodologies for mapping existing XML or other data to the scheme, including the development of procedures and tools
- aiming to establish a union catalogue of all collections based on this work.

The participants agreed that they would be willing, as far as resources allow, to contribute to projects designed to move this strategy forward.

Richard Gartner/Lorna Hughes
17 June 2011