

## Recent Innovations in Parliamentary Libraries ILFA Pre-Conference 2008, Ottawa

### Digitising Hansard: Putting Two Hundred Years of Parliamentary Debates Online

The UK Parliament is close to completing the digitisation of Hansard, the official report of debates in Parliament. The text files have been placed on the internet for free download in XML, and an experimental database and search interface have been developed using an agile, iterative method. The “beta” search interface is available on the internet and features faceted searching. The source code is also available for download and users can contribute to an open online discussion forum which captures queries and suggestions for improvement.

#### The Official Report: *Hansard*

Direct suppression of reports of the proceedings in Parliament in England ceased in 1771 and in the following years there were various attempts to report on Parliament in a systematic manner. William Cobbett began publishing the *Parliamentary Debates* as a supplement to his *Political Register* in 1803, and from 1809 these were printed by TC Hansard. Cobbett sold his interest in the *Debates* to Hansard in 1812 and from 1829 the name *Hansard* appeared on the title page of each issue. The publication was initially based on reprints of reports of speeches from the press, but checked with the Member. The accuracy of reporting in the mid-Victorian *Hansards* has been questioned.<sup>1</sup> In 1877-8 Hansard employed its own reporters and from 1889 the House decided to subsidise *Hansard* so that a permanent record was available. Agreement was reached on more extensive coverage of the proceedings. Hansard was taken into direct ownership by the House in 1909.

#### The Digitisation Project

This joint project of the House of Commons and House of Lords Libraries started in 2005 with a procurement exercise, although there had been a number of exploratory projects in the preceding years. These had not progressed because the costs were too high, the quality of the product was poor, or both.

The reasons for digitising were numerous, apart from the obvious advantages of widening access and enabling free-text searching. Parliament had an unknown number of sets of Hansard, taken from different editions and held in the Libraries of each House and in other departments. Each set occupied approximately 150 metres of shelving. Large portions were printed on poor quality acid paper and had started to deteriorate. They were in regular use, which was increasing the

---

<sup>1</sup> Hansard's Hazards: An Illustration from Recent Interpretations of Married Women's Property Law and the 1857 Divorce Act, Olive Anderson, *The English Historical Review*, Vol. 112, No. 449 (Nov. 1997), pp. 1202-1215

speed at which they deteriorated. Conservation was unattractive as an option as it is costly and does not bring any of the additional benefits of digital capture. Digitising Hansard, on the other hand, would enable the preservation of a single set in a controlled environment and the disposal of some of the remaining sets. It would also provide a reference set of images which could be used to produce facsimilies if needed.

The procurement started with a very loose set of requirements, designed to attract a range of imaginative solutions. The tenders received offered a vast range of prices and specifications but the chosen supplier, AEL Data of Chennai, India, proposed a very high degree of accuracy, impressive generation of metadata and an affordable price. The contract specifies text conversion to an accuracy level of 99.5%.

### **The Digitisation Process**

Scanning is done from a disbound set in order to reduce costs. The disbound pages are discarded when the scanning is complete. The page images are captured as TIFF files at a resolution of 300 dots per inch. Copies of each page are made as JPEG files: this compressed format is more suitable for display on the internet. Optical character recognition (OCR) is carried out using an innovative triple-compare process developed by the contractors. It is similar to the method used for double and triple re-keying of text: three versions of the text are produced from the images in extensible mark-up language (XML) using different OCR software. Discrepancies are identified and corrected and a composite is generated.

Metadata is generated automatically according to a schema, based on the layout of the original text. Semantic as well typographical metadata is captured. For example, in all but the earlier volumes the name of the MP who is speaking is laid out in bold at the beginning of the speech. This and other elements are captured as part of the schema, to enable searching or re-ordering of content with reference to one or more metadata elements. The composite is proof-read in full by the contractors and extensive manual checking of the metadata capture is also carried out.

In addition to the quality assurance work carried out by the contractors, staff of the House of Commons Library carry out further checks. All images are checked for alignment, definition etc using very large screens and specialist software which enables multiple image display. Other QA processes, such as checking for common text conversion errors, are automated. Higher levels of the schema are checked and finally samples of the text are proof read. This has confirmed that the high contractual level of accuracy has been met and in most cases exceeded. Metadata capture has also been very successful, but in most files the schema does not validate so some remedial work on the text has been necessary.

The complete run of Hansards from 1803 to the general election of 2005 has been captured. This constitutes about 2.75 million pages. As of May 2008 the digitisation work is substantially complete, except for a relatively small number of volumes which require re-scanning. The period of 1988 to 2005, for which an online version is already available, was included to improve the completeness of the data set and to take advantage of the more detailed metadata which is created as part of the digitisation process. Digitisation of the debates of the Standing Committees (the former name of the committees which carry out the “committee stage” scrutiny of Bills in the House of Commons) will start shortly.

### **Use of the Digitised Content**

The XML text files which are available to date have been posted on the internet and are available for download via the Archives section of the UK parliament website (<http://www.parliament.uk/publications/archives.cfm>). A downloadable “click use” licence is available for free onward use of the text, including for commercial use. A prototype database and search interface has been developed by Parliament using open source software.

A small team of developers worked closely with users inside and outside Parliament to develop the search interface. The XML is parsed into HTML before loading it into the database. As of May 2008, about half of the available data has been loaded, taken mainly from the 20<sup>th</sup> century. Browsing and searching of the text can be accomplished using a number of “facets”. This allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in a variety of ways, rather than in a single, pre-determined order. The facets used include date, volume number, monarch, chamber, content type (debates or questions), constituencies (electoral divisions), Members of Parliament and offices held.

Navigation by date can be approached by drilling down using bar chart-type graphics representing decades, years, months and days. A faceted approach to search enables filtering of search results using (for example) date, Member of Parliament or content type. References in the standard format using date, volume and column numbers (eg HC Deb 13 May 2002 Vol 385 c498) can be located using the search box. Pages for individual Members of Parliament have been created, giving details of constituencies represented, first and last speeches made and total number of contributions recorded. Constituency pages have been created, listing all Members which have represented them.

We are currently working to reconcile the metadata on entities such as Members and constituencies with authority lists to improve accuracy. For example, inconsistencies in name format in the original text (John Smith, Mr Smith, Mr John Smith etc) can produce multiple entries for a single entity. References to Acts of Parliament and Bills were not tagged as part of the schema but these can

now be identified automatically and hyperlinked to individual pages for each Act and Bill. Division lists (the record of votes taken by each House) are tagged and we are working on making these available in a machine-readable format so that they can be easily downloaded for statistical analysis. We have developed a simple geographical interface, plotting references to place names in Hansard for a given day on a Google map and linking through to the text of the debate. We are now working on a more useful geographical interface which will enable the identification of alternate place names based on proximity to the original geographical search term.

A public version of the prototype was made available in December 2007 (<http://hansard.millbanksystems.com>). Since we are still experimenting, not all of the features described above will be available on the public version at a given time, and performance is variable if new data is being loaded. The latest version was launched in late May 2008. Feedback on the prototype is welcome via a discussion group, a link to which is located at the contact page (<http://hansard.millbanksystems.com/contact>). An open issues log, describing problems to be resolved and future developments, is maintained in the discussion group area.

The JPEG page images have not been utilised in the prototype since user feedback has not indicated a strong desire to see images of the original text. They may be used in the future and will be made available to third parties who may wish to host the Hansards themselves. The TIFF images will provide a means of producing facsimiles of the original volumes or re-running the OCR process should this be required in future. The files represent about 40TB of data and are currently stored on portable hard drives. Preservation of the electronic files for the indefinite future is being investigated. It is ironic that a project which aims to assist the preservation of 19<sup>th</sup> and 20<sup>th</sup> century records should highlight the emerging issue of digital preservation!

## **The Effects of the Project**

A number of aspects of this project have been innovative:

- Digitisation of large volumes of parliamentary proceedings
- Use of overseas contractors
- Innovative digitisation methods leading to a high quality product
- Availability of the entire text as a free download for onward use
- Use of open source technology to develop a prototype “front end” for the content
- Use of agile, open, highly iterative development methods for the front end, allowing a high degree of input from users
- Use of faceted classification and search methods for retrieving the content

The full impact of the project has yet to be felt as the front end is still in a “beta” version, is not linked to the official parliament website and has not been widely publicised. Even in its beta version the site is receiving over 1500 hits per day, mainly through Google.

The principal benefit will be to allow free online access to Hansard for the public in the UK and the rest of the world, using an intuitive but highly sophisticated search interface. A benefit already achieved as a result of the development methods employed is that internal and external users have a high degree of confidence in the product and have offered many valuable suggestions for improvements. In addition, the search interface has introduced many stakeholders within Parliament to the concept of faceted classification, and is proving influential for other projects such as the development of new search tools for the official parliament site.

Edward Wood  
Director of Information Management  
House of Commons  
United Kingdom  
28 May 2008