Structuring Hansard transcripts into interoperable and reusable data

Rachel Barber-Pin and Yasuko Enosawa[1,†]

[1] *Legislative Assembly of Ontario, 111 Wellesley Street West, Toronto, Ontario, M7A 1A2, Canada*

Abstract

Information Services Branch at the Legislative Assembly of Ontario began exploring better methods of how to improve access to legislative information frequently requested by MPPs, their staff and the public. As open, structured content is the most empowering development of legislative information dissemination since the Legislative Assembly began publishing content online, our focus was to turn documents into usable data. This presentation will be a case study of how we structured the House Hansard transcripts to become interoperable and reusable data. It will provide a summary of the Advanced Hansard Search (AHS) Modernization Project and examples of how the structured Hansard transcript datasets are being used in different applications and services at the Legislative Assembly.

Keywords

Metadata interoperability, parliamentary debates, Hansard, structured documents

1. Advanced Hansard Search Modernization Project

The House Hansard is an official report of debates and proceedings in the Legislative Assembly and has a historically significant role in informing the public. House Hansard transcripts are digitally available from March 1974 to the present on the Assembly's website. The Advanced Hansard Search tool was available through the website, but the search function was not intuitive, and it was built on unsupported technology.

To modernize the search tool and improve the search experience, the Advanced Hansard Search (AHS) Modernization Project was conducted from October 2023 to March 2024. Drupal, a website content management system, was used for building a user interface. Solr, an open-source platform, was used as a backend search server. ParseHub, Excel and OpenRefine were used for content scraping, cleaning and structuring.

To create structured Hansard datasets, the core metadata elements ware applied to the data: Parliament-Session numbers, Date of the debate, Type of Business, Topic, Name of person speaking, and Intervention (spoken content).

2. Challenges

At the beginning of the AHS project, a Hansard transcript was limited to three formats: Word, PDF, and flat HTML. These formats had limited machine readability and staff were

[†] These authors contributed equally.

rbarber-pin@ola.org (R. Barber-Pin); yenosawa@ola.org (Y. Enosawa)

https://doi.org/10.23106/dcmi.952452720

unable to pull out datasets across all transcripts. The original Hansard Search tool was created in the 1990s, and the original schema no longer met the current needs of the new platform. A new schema was required to develop for structuring transcripts to support a new tool that would meet the search requirements of internal staff and external users.

There were also issues with transcripts themselves. The first issue was that transcripts have not been consistently formatted over the years, so the team needed to use various scripts to scrape the data from the HTML format into the Excel files with the set schema to ensure that the data could be used for the new search tool. The second issue was that transcripts contained inconsistent terminology which led to issues when structuring and cleaning the data. The team had to manually go through and clean any data that had not been captured accurately in the scrapping process.

3. Outcomes

The metadata schema used for the AHS project led to the outcomes of three information products/services: Visualized Hansard, Hansard Search, and shareable open datasets. All of them use the structured Hansard Search transcript datasets which allow for greater resource discovery and availability. The Hansard Search tool uses the datasets that populate the Solr platform which uses a defined schema that identifies: type of field acceptance, storage handling and field type index and querying. Visualized Hansard is a digital tool to help users discover transcripts in a visualized user experience. Users can filter paragraphs of what was said in the Legislative Assembly by member, order of business, and subject of business. Finally, the datasets created from the project were also published online on the Assembly's website in CSV format as open datasets for users to access, and reuse.

4. Enhancements

Now that the redesign of the Hansard Search tool has been completed, Assembly staff are working with stakeholders to plan for an enhancement project.

One enhancement activity that we will be exploring is adapting the schema to allow for a broader level of speech intervention by a member. Currently the schema has a "paraurl" element which allows detailed searches by paragraph, but this can unnecessarily bloat search results. Assembly staff are now researching how to incorporate an element that allows for the structured data to be searched by a member's whole speech, rather than just at the paragraph level.

Research from the AHS Project will be applied to other potential projects such as the Committee Hansard search tool. Assembly staff aim to look at how the House Hansard schema could be leveraged for the Committee Hansard and whether both the House and Committee schemas could be mapped together in one search tool.

https://doi.org/10.23106/dcmi.952452720