Diseño y Procesamiento de Documentos XML

Trabajo Práctico Especial

Gorostiaga, Pablo - pgorostiaga@itba.edu.ar

Pampuri, Franco Manuel - fpampuri@itba.edu.ar

Glaubart, Jonas - jglaubart@itba.edu.ar

Migliaro, Eugenio - emigliaro@itba.edu.ar

GRUPO 23

2024 - 2C

Introducción:

Este trabajo implementa un sistema de procesamiento de datos para consultar y visualizar información histórica del Congreso de los Estados Unidos a través de la API oficial de Congress.gov. El sistema permite acceder y procesar datos estructurados sobre la composición y actividad de los congresos históricos, incluyendo detalles sobre sus cámaras legislativas (Representantes y Senado), sus miembros, y las sesiones realizadas. Todo esto mediante herramientas como XQuery, XSLT y Bash. El objetivo principal es desarrollar un proceso automatizado que, a partir de un número de congreso específico (entre 1 y 118) genere una visualización estructurada en formato HTML, de información específica de ese congreso.

Desarrollo:

1. Trabajo realizado:

El programa sigue tres etapas:

I. Adquisición de Datos:

El proceso comienza con un script de shell (‘tpe.sh’) que acepta un número de congreso (1-118) como entrada y realiza una validación inicial. Usando la API de Congress.gov, realiza dos peticiones HTTP: una para obtener información general del congreso y otra para recuperar detalles de los miembros. Estas solicitudes requieren una clave API almacenada en la variable de entorno CONGRESS_API. Los datos se guardan en formato XML como ‘congress_info.xml’ y ‘congress_members_info.xml’. El script bash valida los valores ingresados y retornados y maneja los posibles errores de acuerdo a un archivo errors.xml que relaciona cada código de error posible con su descripción. Finalmente pasa como parámetro el código de error al comando XQuery.

II. Transformación de Datos:

Un script XQuery (‘extract_congress_data.xq’) procesa los archivos XML sin procesar para crear una estructura de datos estandarizada. El script une y transforma los datos de ambas fuentes, organizándose jerárquicamente por cámaras (Cámara de Representantes/Senado), con información detallada sobre miembros y sesiones. La transformación incluye la normalización de campos de texto, ordenamiento de miembros por nombre y estructuración de información

temporal sobre períodos y sesiones del congreso. La salida se valida contra un Schema XML que define tipos estrictos para todos los elementos, incluyendo tipos complejos para información del congreso, cámara, miembros y sesiones. Los datos transformados se guardan como 'congress_data.xml'.

En caso de haber sucedido un error en la adquisición de datos, el archivo xml generado sólo contiene información sobre dicho error.

III. Presentación:

Finalmente, una hoja de estilos XSLT ('generate_html.xsl') convierte los datos XML estructurados en una presentación HTML. La hoja de estilos crea una visualización jerárquica con el nombre del congreso y el período en la parte superior, seguido de secciones separadas para cada cámara. Dentro de cada sección de cámara, genera dos tablas: una para miembros (mostrando nombre, estado, partido y período) y otra para sesiones (mostrando número, tipo y período). Las tablas están formateadas e incluyen capacidades de ordenamiento para los nombres de los miembros. La salida final se guarda como 'congress_page.html'.

2. Dificultades:

En cuanto a dificultades, estas fueron pocas pero bastante específicas. Nuestro mayor problema fue recordar el uso de funciones y sintaxis de bash, lenguaje que aprendimos en la materia 'Introducción a la Informática'. Esto no nos afectó en términos de no poder avanzar con el proyecto, pero sí hizo que sea más difícil poder poner en código nuestras ideas. Otra complicación que tuvimos fue el proceso de verificar que el usuario intentando usar el programa tenga definida la variable de entorno.

Además, tuvimos ciertas complicaciones con la información que extraemos de la página, la cual tenía espacios y tabuladores extras, lo cual nos generaba errores al momento de comparar. Esto nos llevó al uso de normalize-space, que mencionaremos más adelante.

Y por último, tuvimos problemas con el acceso a la información específica de cada mandato, de cada miembro del congreso. Nuestra dificultad estuvo en que en algunos pocos casos, ciertos miembros estuvieron presentes en la misma cámara durante dos periodos de tiempo

distinto, o tuvieron un mandato en una cámara, y un mandato en la otra cámara. Esto nos trajo problemas, ya que al momento de acceder a la información de los miembros en estas situaciones, nos encontrábamos con que la información adquirida contenía una unión de dos o más strings, y esto nos traía errores al momento de hacer comparaciones con esta información. Esto nos generó bastantes problemas y fue una gran dificultad el poder resolverlo.

3. Investigación extra:

Como ya se mencionó anteriormente, investigamos acerca del uso de normalize-space que fue esencial para poder resolver algunos de los problemas que tuvimos. Otro concepto que investigamos fue la función not, la cual nos ayuda a verificar que los argumentos no sean vacíos. Un aspecto que tuvimos que investigar en detalle fue la validación de parámetros cuando se trataba con la API. Tuvimos que investigar como poder hacer para verificar que la API esté correctamente configurada en el entorno.

4. Roles de los Integrantes:

Aunque no definimos roles específicos en el grupo, es posible identificar una división de tareas según las contribuciones de cada integrante:

1. Funcionamiento de la consulta XQuery: Franco Manuel Pampuri
2. Funcionamiento de la plantilla XSLT: Eugenio Migliaro
3. Funcionamiento global del proyecto: Jonás Glaubart
4. Presentación: Pablo Gorostiaga

Conclusiones:

A lo largo del desarrollo de este trabajo se pudieron integrar y aplicar la gran mayoría de los conocimientos asimilados durante la cursada de la materia. Adicionalmente, gracias a ciertas dificultades u obstáculos que se presentaron, se logró incorporar herramientas nuevas que permitieron mitigarlos.

En definitiva, la cualidad práctica del trabajo dio lugar a un enfoque distinto a tomar para su resolución en comparación con otras actividades de la materia, esto facilitó no sólo la interpretación de todo lo aprendido, sino también la profundización sobre dichos conceptos.