Mohammed Ajmal
E-mail: ajmal@comm.toronto.edu

PPTX Parser

Overview

We demonstrate a simple tool which parses Microsoft PowerPoint 2007 presentations and converts them into a custom XML representation. The parser is able to handle:

- Bulleted lists (arbitrary nesting & depth)
- JPEG images (including Exif metadata)
- Tables

In addition, we include a XSL stylesheet which allows users to view the custom XML representation of a presentation in HTML format. Finally, we show proof-of-concept security features which enable users to block keywords within presentations.

Details

A high-level diagram for the entire system is shown below, in Figure 1. As indicated, the PPTX parser takes as input a slide presentation in OpenXML format, the new format adopted by Microsoft for all the components of the Office suite. Note that this is one of the key reasons we have chosen to operate with Microsoft PowerPoint 2007 presentations. The OpenXML format is an open-source format. In addition, the use of XML technology to represent presentations allows great flexibility in terms of our ability to support new features that Microsoft may release at a later date. Finally, the overwhelming majority of presentation slides are authored using Microsoft PowerPoint. Consequently, by operating with this format, we are able to target the largest set of users.

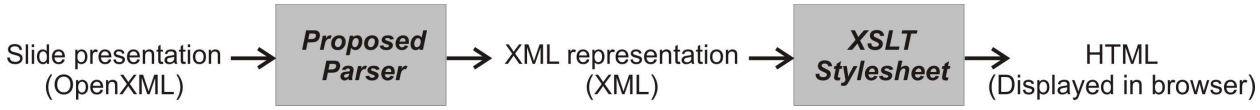


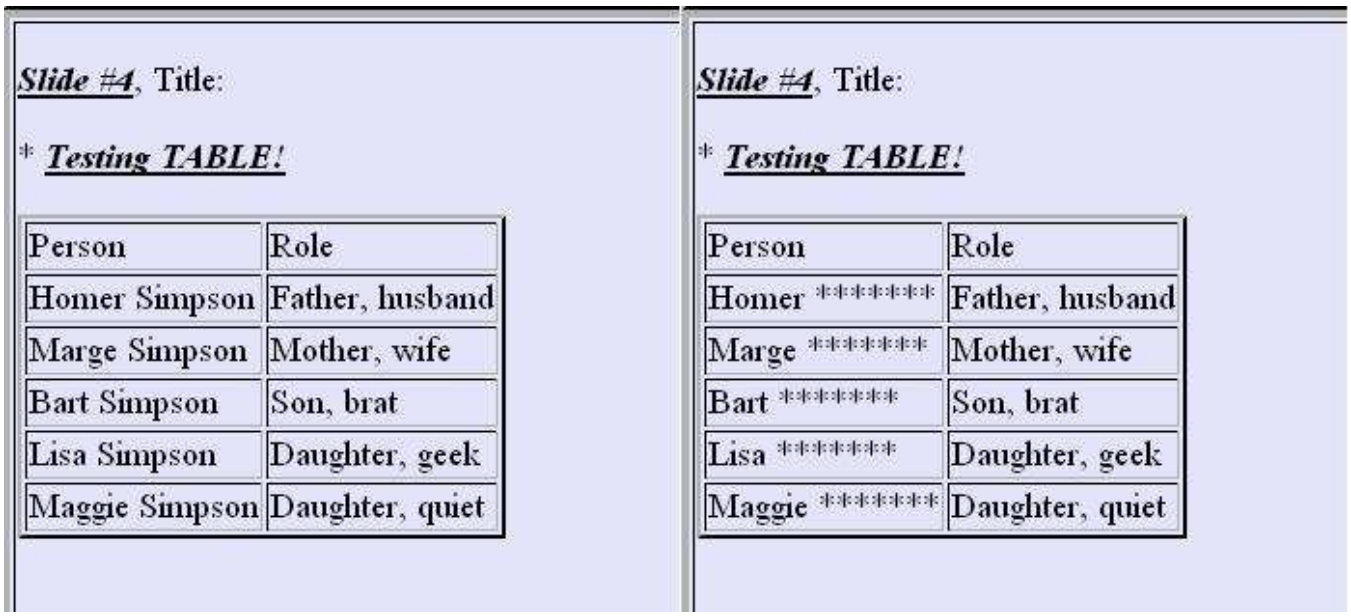
Figure 1: System diagram.

Note that the output of the PPTX parser is a representation of the original presentation in a custom XML format. This serves two main purposes: first, we have a more concise representation of the *relevant* information from the slides. Secondly, the parser has attached semantic meaning to each of the elements of the presentation. For example, we extract EXIF metadata from JPEG images in the presentation which allows for semantic processing of images. As a result, the custom XML representation is useful for performing further tasks on the presentations (ex: summarization, slide retrieval).

Finally, after successfully parsing the slides and assembling the data, the last stage in the system presents the information in a suitable manner. Since our representation of the parsed data is also an XML file, it is extremely simple to display the presentation in any format we desire. Indeed, this is another advantage of the custom XML representation. For this demonstration, we have chosen to render the presentation in a web-browser, using standard HTML. It should be noted, however, that there is no difficulty in having the final stage output a modified PowerPoint presentation, or any other format. We use an XSL stylesheet to perform the conversion from the custom XML to HTML.

Security

We have added simple functionality that allows the user to specify keywords that should be blocked from the slides in a presentation. When a user specifies a keyword that they wish to omit from the slides, all occurrences of the word in the slides are replaced by a series of asterisks. Users also have the option of unblocking keywords. This functionality is shown in Figure 2. The table on the left lists the core members of the Simpsons. The table on the right is the resulting table when the keyword 'Simpson' has been blocked.



Slide #4, Title:

* ***Testing TABLE!***

Person	Role
Homer Simpson	Father, husband
Marge Simpson	Mother, wife
Bart Simpson	Son, brat
Lisa Simpson	Daughter, geek
Maggie Simpson	Daughter, quiet

Slide #4, Title:

* ***Testing TABLE!***

Person	Role
Homer ********	Father, husband
Marge ********	Mother, wife
Bart ********	Son, brat
Lisa ********	Daughter, geek
Maggie ********	Daughter, quiet

Figure 2: Blocking user-specified keywords.