Task # 1428: האילמה ילוקוטורפ רוזחא - החותפ תסנכ

Status:	New	**Priority:**	Normal
Author:	בירק סדא	**Category:**	סינותנ תריצק
Created:	10/26/2010	**Assignee:**	Ori Hoch
Updated:	01/21/2013	**Due Date::**	
הנושאר הלטמכ סיאתמ:	No		
Description:	הדעוו לכ סע סויה השענש המל המודב		

History

12/24/2012 07:29 pm - Ori Hoch

- Assignee set to Ori Hoch

- Category set to סינותנ תריצק

- הנושאר הלטמכ סיאתמ set to No

need to get the protocols from this url:
http://www.knesset.gov.il/plenum/heb/plenum_search.aspx
and then parse the doc files

12/29/2012 05:35 pm - Ofir Carny

- File parse.py added

Attached script contains python retrieval code, openoffice based parsing and HTML conversion, and some post conversion HTML parsing and Hebrew unicode handling

01/07/2013 10:08 pm - Ori Hoch

- File 03581012.doc.awdb.xml added

- File 04085512.doc.awdb.xml added

- File 18_ptm_219371.doc.awdb.xml added

attached: sample xml files produced using antiword

wiki page for planning the parsing process of these xml files:
[[https://github.com/astupidog/Open-Knesset/wiki/parse-plenum-protocols]]

01/13/2013 03:38 pm - Ori Hoch

- File plenum_htmls.zip added

parsing process is finished and working but there are probably many bugs..

attached are html files that were parsed, hopefully someone will look over them and provide some bug reports

the plenum_htmls.zip file should be extracted somewhere, then you should open the index.html file using a web browser (tested with chrome but should

work on other browsers as well)

01/21/2013 08:45 pm - Ori Hoch

רשמהל תודוקנ:

1. קר היהת ."האילמ תדעו"כ וא "הדעו"כ התוא ומסיש הדעו לכל "type" הדש ופסוותי רשאכ תודעו לש תואלבטו מילדומ מתואב שמתשנ מינותנה דסמב מילדומ תניחבמ תחא "האילמ תדעו".
2. הנוש parsing עצבת זאו "האילמ תדעו"ב רבודמ מא קודבתש רכ create_protocol_parts לש היצקנופל יאנת תפסוה.
3. וכו "רבוד אלל טסקט", "תרתוכ", "רבוד" :גוס לש הדש ופסוותי שאכ protocolParts לש ודבועי מיקלחה לכ'.
4. תסנכה רתאמ לוקוטורפה תא בוש דירוהל ילב re-parse עצבל לכונש רכ לוקוטורפה לש יקנה טסקטה תא רומשל שמשי protocol_text הדשה.

Files

parse.py	7.8 kB	12/29/2012	Ofir Carny
18_ptm_219371.doc.awdb.xml	459.9 kB	01/07/2013	Ori Hoch
03581012.doc.awdb.xml	414.8 kB	01/07/2013	Ori Hoch
04085512.doc.awdb.xml	249.7 kB	01/07/2013	Ori Hoch
plenum_htmls.zip	2.9 MB	01/13/2013	Ori Hoch