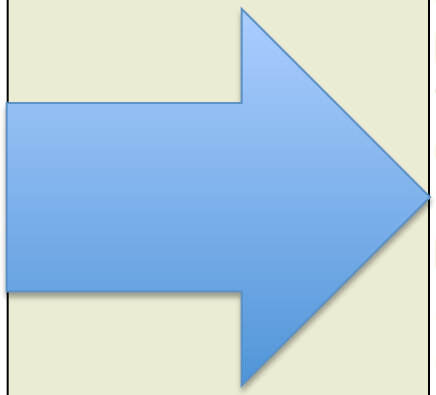
Tabularization in PanLex

David Kamholz

PanLex Summer Internship

27 June 02013

```
<tr>
<td>man</td>

<td>lac</td>

<td>lalaki</td>

<td>balances</td>

<td>tigban</td>

<td>timbangan</td>
</tr>

<tr>
<td>woman</td>

<td>paranpaon</td>

<td>babaye</td>

<td>weight</td>

<td>tahil</td>

<td>timbang</td>
</tr>
```



```
man lac lalaki
balances    tigban  timbangan
woman   paranpaon   babaye
weight  tahil   timbang
young woman beni beni   dalaga
pearl   mutiara mutya
married woman   babay   babayeng minyo
mother of pearl tipay   tipay
buttocks    samput  balat-ang
pipe (musical instrument)   subing  plawta
thigh   paha    paa
disease of St. Job  alupalan    hubag
knee    tuhud   tuhod
bring me    palatin comorica    ambi
shin    bassag-bassag   bitiis
certain rice cakes  tinapai puto‧bibingka
ankle   bolbol  buulbuul
good    main    maayo
heel    tiochid kiting
no  tifale  dili
sole (foot) lapa lapa   lapalapa
knife   capol‧sundan    kutsilyo
gold    balaoan bulawan
scissors    catle   gunting
silver  pilla   plata‧pilak
(to) shave  chunthinch  mamalbas
brass   concach tumbaga‧bronse
```

Tabularization strategies

- Automatic
  - write scripts to do everything (in Perl or some other language)
- Semi-automatic
  - perform a few manual tasks first, e.g., eliminating irrelevant content; copying and pasting document portions together; manually removing irregularities
  - write scripts to do the rest
- Manual
  - re-type the text of an image, or reformat an irregularly structured document

Digital file formats

- Text
  - large diversity of formats, some standard (Toolbox/ MDF, .wb)
- HTML/XML
  - large diversity of formats
  - tag syntax and hierarchy is standard (but many web pages do not have well-formed HTML)
- PDF
  - major categories: text, OCR-able scans, un-OCR-able scans

Text tabularization strategies

- Most general strategy is to read the file line by line and parse it with regular expressions
  - works very well for simple cases
  - can work for more complex cases, with some effort and ingenuity
- For more highly structured formats (e.g., Toolbox/MDF), it is generally necessary to use something other than line breaks to divide data into records
  - one technique: read the whole file in at once, then *split* it on the relevant delimiter, if there is one
  - another technique: implement a stateful parser that keeps track of what part of a record it is in on each line

HTML tabularization strategies

- Three major strategies
  - treat the raw HTML as a text file
  - extract the text (e.g., using copy and paste from a browser), and tabularize the output
  - use a library to parse the HTML into a hierarchy of elements, and extract the tabular data from the parsed result
- Best strategy depends on the file
- See http://dev.panlex.org/tools/ for suggested HTML parsing libraries

PDF tabularization strategies

- First extract the text, then tabularize as a text file
- Best tools: pdftotext, pdfminer, export from PDF readers, copy and paste from PDF readers
- Important to specify UTF-8 as output encoding
- Challenges
  - some PDFs have custom fonts or encodings and will require manual correction
  - difficult to faithfully preserve details of text layout and formatting
    - pdftotext has *-layout* option
    - pdfminer has HTML output format
  - difficult to unwrap columns
    - pdftotext has *-raw* option

Tabularization in phases

- Sometimes it is useful to tabularize a file in multiple phases
- Example: a text file with records that wrap to multiple lines, with a clear indication of whether any given line begins or continues a record
  - phase 1: unwrap multiline records so that they are on single lines
  - phase 2: tabularize the output of phase 1