GSOC 2009: MediaWiki export formats

There are more than one way to export MediaWiki content. In this page I will describe the methods I studied as possibilities to use in the MediaWiki - TikiWiki importer.

MediaWiki XML export feature

MediaWiki has a built-in XML feature to export all wiki page content. It does not export users and the XML contains the MediaWiki syntax (no wiki syntax parsing is done).

It has a easy to use command line script called dumpBackup.php that output all wiki pages with history but also accept a lot of different arguments to export only the last version of each page and so on.

MediaWiki XML Bridge

MediaWiki XML Bridge extension is another tool to export wiki pages to XML format (or in this case also XHTML). It uses the mwlib, a python library to parse MediaWiki articles.

Nelson question to help evaluate XML Bridge:

1. Is XML Bridge any good?
   Rodrigo: I'm not confident that XML Bridge is something interesting for our project. Apparently they use a non standard and MediaWiki specific XML representation called mwxml. I wasn't able to find the format specification. Also, mwlib is oriented to fetch through HTTP only the last revision of an article. mwlib is developed by pediapress.com, they print books from MediaWiki sites. Maybe that is why they are not concerned with wiki page history.
   As XML Bridge doesn't export the page history I don't think it might be useful for the MediaWiki to TikiWiki importer.
2. What should we write to convert this XML to Tiki? (maybe we can write a PHP XML bridge in reverse to Tiki or maybe stick with Python)
   Rodrigo: A mwxml parser 😞
3. Is the XML representation a standard to wiki conversion?
   Rodrigo: No, XML Bridge use mwxml a XML representation specific for the MediaWiki syntax. I wasn't able to find the format specification.
4. Is XML Bridge to MW a two way bridge? I suppose it is. Is it lossy? Are some syntax lost?
   Rodrigo: I'm not sure if XML Bridge is two way. I didn't found in the documentation any way to insert content in a wiki page using the mwxml format. Also, I didn't found any reference to be sure if mwxml support 100% of the MediaWiki syntax or if there is syntax loss. Probably there no significant syntax loss as XML Bridge uses mwlib which is the official way supported by the MediaWiki foundation to export MediaWiki articles to formats such as PDF or OpenDocument.