Character substitutions

This is the coordination page for: Character substitutions in Tiki

The code started in r43471.

Background

In Tiki, page names are case insensitive. "Commit Code", "commit code" and "COMMIT CODE" are all equivalent.
sylvieg: Are you sure? - in mysql as case insensitive - but in postgres, I do not think so. I am enable to find in the code a place where the strlower is done
Jyhem: I believe Tiki does not enforce this. Your database does. If I create a mySQL database with "utf8_unicode_ci", it is like Marc describes. When I create the database with "utf8_bin", then it is not true: "test" and "TEST" are different pages.
This makes things simpler for end users and has never, AFAIK, been reported as an issue on the wishlist. There is also a substitution for Microsoft Word Special characters.
sylvieg: so far I found in the case there is an utf8 normilization, nothing else
alain_desilets: I think character substitutions makes sense for latin languages. But what do we do with Chinese for example?

What about accents? spaces? underscores? plus? slash? etc.

Do we really want a page "Déjà vu" and a different page "Deja vu"? Probably not. Thus, we should identify characters which would serve as aliases. Let's determine this together.

Since wiki page names should avoid special characters, we should consider using character substitutions in page names (a instead of à, etc.) and use the description field for the exact format. Using the description is also useful for very long wiki page names.

Objectives

- Easy for the end user
- Clean URLs (avoid %20, and similar)
- Robust
- Handle cases where there is a desire for similar but different page names, where current behavior is a feature, not a usability bug.

Questions

- How does Wikipedia do it?
    - http://en.wikipedia.org/wiki/Help:Page_name
- Should it be aliases, redirect or substitutions?
    - Aliases: all work
    - Redirect: all work, but you are redirected to the cleaner URL, for nicer copy-pasting
    - Substitutions: when you create a page with a special character, Tiki swaps for another character.
- What is universally accepted in URLs? (without conversion to %20 or similar) (This is standardized in RFC 3986)

- What about languages like Arabic and Mandarin?

Suggested alias/substitutions

Character	Could/Should be	**But**
a-z and A-Z	No special handling	
parenthesis (( ) or ( ))	No special handling	
All characters with diacritics (à, ç, é, î...)	Equivalent without diacritics (a, c, e, i...)	I see a problem ignoring accents there ! For example in Czech language: "Most" and "mošt" are two very different things (a bridge and a cider) Just my 2 Czech crowns... — luci. French (and presumably most languages with diacritics) are the same. — Chealer
Space ( )	hyphen (-)	
Plus (+)	hyphen (-)	This is problematic. If I create a "Visual C++" page, I do not want it named "Visual C--". And **this kind of issue probably applies to all substitutions**.
Apostrophe (')	hyphen (-)	
Colon (:)	hyphen (-)	
SemiColon (;)	hyphen (-)	
Slash (/)	hyphen (-)	
Backslash (\)	hyphen (-)	
Pipe (|)		
Ampersand	(&)	So no substitution?
At (@)	should find an email address if we just search for prefix or suffix	What?
number sign	(#)	So no substitution?

Discussion on #wiki about # in pagenames

[+]

Other comments & questions

sylvieg: and what about in a first step, work on the like pages that are proposed when editing a page. The like pages proposition is very poor , I think for the moment - it is only 'contains a word in common' - why not keeping a normalized form of the pagename in the database (obtained by a replacement pattern defined in admin)...

Question?
Should hypen (-) and underscore (_) be aliases?
comma (.) -> conflicts with ShortURLs, yet, it's common to want a page name with one. What to do?
What about dollar ($) sign?

Anchors in automatically generated table of contents

For example: http://dev.tiki.org/EditUIRevamp#Preview_amp_history

Usernames

Also, usernames should follow similar, if not the same guidelines.
Sylvie: username has a filter hardcoded in 2.x - now a param in 3.0

admin->username pattern - the default is

```
/^[ '\-_a-zA-Z0-9@\.]*$/
```

Gmail prevents certain characters to avoid confusion between two users. Perhaps we should do the same?

In Facebook and Gmail, joesmith is the same user as joe.smith

Search engine

Maybe this should be used for search engine results as well. Searching "Déjà" should find "Deja" and vice-versa.
When I search "event", I would like to find this page: http://profiles.tiki.org/Event_Management_System

Tags

Tags can already get messy with synonyms, plurals, typos, etc.
Some people use commas (,) in tags entry box instead of spaces and they are recorded.
Sylvie: tag have already a normalisation + lower reduction in settings

Wiki Link Format

Controls recognition of Wiki links using the two parenthesis Wiki link syntax page name.

3 choices in tiki-admin.php?page=wiki are complete, latin and English

Please see related issue reported by cellvia2

Related info

http://en.wikipedia.org/wiki/Punctuation
http://wiki-translation.com/ (This is increasingly important as we improve Tiki i18n features)
http://ca.php.net/strtr (scroll down) pieces of PHP code that rewrite strings into a URL-suitable form.

Related wish list items

- Wiki page names and links, plus (+) simulates a space( ), but a space( ) doesn't simulate a plus (+)
- Quick Edit module should detect and warn about special characters in page names
- period (.) in page names conflicts with Short URLs rewrite rules
- Equivalent characters for page linking, backlinking, searching, etc (ex.: space, underscore, period)

See also

- URL Rewriting Revamp
- Bad characters