Tiki DB Redactor

Code


Idea

Have a cool tool to pass databases around for debugging purposes without disclosing too sensitive information, and to avoid the debugging process to send out watch emails for example to the "users" of the site when it is not real activity. If any emails get sent out it could also contain links to the testing site which confuse users further. Added benefit: db dumps for debugging are small. Some kind of Tiki DB Anonymiser.

Initial use case should be for *.tiki.org content, and later on, this can be improved so it's useful for any Tiki instance.

This should be done with the Tiki Console framework

Problems

It's the worst idea ever, see for example: A Face Is Exposed for AOL Searcher No. 4417749 or Identifying People using Anonymous Social Networking Data.

As every need for redaction stems from another problem, it is impossible to create the perfect tool for all of them. **We don't even know what we have to anonymise:** Fitness tracking app Strava gives away location of secret US army bases.

If the users of a Tiki site do not agree with the passing around of the underlying database dump - whether the redactor is used or not - it is always a misappropriation of community members' data they entrusted to their service provider!

Use cases

- Performance testing: devs need a real-World data set to see where the bottlenecks are
- New feature development.
  - We are working to develop various Natural language processing tools at TikiFest NLP 11 and we need data to develop them on.

Things to redact

Basically **everything** that is not needed for the final use case, but usual suspects that promise to raise the cost of gathering individual-related information are:

user data

- credits and payments (tiki_payment_* ,tiki_credits*, tiki_acct*) **priority high**
- user names (users_users) **priority medium** partly, some other tables still have them
- email, password (users_users) priority high partly, just as user names
- user bookmarks (tiki_user_bookmarks_urls) priority low
- user calendars (tiki_calendars, tiki_minical_events, tiki_minical_topics) priority low
- user contacts (tiki_webmail_contacts) priority low
- user files (tiki_files, tiki_file_drafts, tiki_images) priority low
- user mail accounts (tiki_user_mail_accounts, tiki_mail_queue) priority high
- user messages (messu_messages, messu_archive, messu_sent) priority high
- user notes (tiki_user_notes) priority low
- user tasks (tiki_user_tasks) priority low
- user watches (tiki_user_watches) priority high emails redacted

session data
- sessions priority high
- tiki_cookies priority high
- tiki_sessions priority high

tables containing ip addresses / email addresses
- tiki_actionlog priority low
- tiki_banners priority low
- tiki_banning (ip addresses) priority low
- tiki_invited (email) priority high
- tiki_newsletter_subscriptions (email) priority high
- tiki_sent_newsletter_errors (email) priority high
- tiki_logs (username / ip matching) priority low
- users_users (email) priority high

tables containing passwords
- tiki_dsn (db-passwords) priority high
- tiki_mailin_accounts priority high

global tiki configuration data
- google connection data (map api key, ...) priority high
- intertiki config priority high
- ldap connection data etc. priority high
- login passcode if it's sent by admin only priority high (what's this?)
- a variety of access tokens/api tokens for third party apps. priority high
- register passcode

other tables with general privacy problems on export
- tiki_auth_tokens (auth-tokens, email adresses) priority high
- tiki_connect? priority medium
• tiki_forum_reads (general privacy issue) priority low
• tiki_history (mixed junk of old versions of public and private items of all kind) priority low
• tiki_live_support_messages (may contain emails and passwords) priority low
• tiki_live_support_requests (may contain emails and passwords) priority low
• tiki_mail_events (email addresses) priority high
• tiki_preferences priority low
• tiki_referer_stats priority low
• tiki_source_auth priority low
• tiki_user_reports_cache ? priority low
• tiki_webservice (may contain private urls and login data for webservices) priority high

strip tables to make the archive smaller
• tiki_secdb priority low
• tiki_history priority low
• caches for urls etc. priority low

*.tiki.org specials
• user data in trackers priority medium

more things
http://sourceforge.net/p/tikiwiki/code/47257

Comments:
• emails: even better is to have an option to replace by test mails priority medium
• objects: remove all wiki pages, blog posts, tracker items, files, etc. not visible to anonymous users (so keep data that could be crawled) priority low

Future Ideas
• Connect & Redactor: use a pref attribute instead of maintaining lists of prefs

Related links
• https://github.com/fzaninotto/Faker

alias
• Tiki DB Scrubber
• Tiki DB Anonymiser
• Redactor