Tamil Wikipedia – Google Translation Project

Overview and Lessons learnt

Bala Jeyaraman
User ID : Sodabottle

Tamil Wikipedia – Google Translation Project

Tamil Wikipedia

- ~ 80 million speakers
- Native populations in India, Sri Lanka, Singapore, Malaysia
- Significant diaspora presence – North America, Europe, Australia, South Africa and Mauritius
- ~39,500 articles (October 2011)
- ~20 editors with >100 edits a month

Tamil Wikipedia –
Google Translation Project

Google:

- DO NO EVIL (or so we believed)

Tamil Wikipedia – Google Translation Project

Overview:

- English wiki to Tamil wiki
- Google Translate Toolkit (now released as alpha version)
- June 2009 – March 2011
- Translators - Third party vendors – hired by Google
- ~ 1250 articles created
- Project ended in failure

Tamil Wikipedia – Google Translation Project

Time line:

- June 2009 – First Machine translated articles started appearing. No intimation to the Tamil wiki community

- Aug 2009 – No of articles increase - No response to community queries

- Oct-Nov 2009 – Accidental reveal of project existence; Concerns increase

- December 20, 2009 – First official meeting with Google .

- Jan 2010 – Community considers scrapping the project; decision taken to formulate guidelines and communicate with Google

Tamil Wikipedia – Google Translation Project

Time line:

- Feb 2010 – Request for Comment; community guidelines formed
- Feb 2010– Coordinators appointed
- March 2010 - Required article list handed over
- July 2010 – Second meeting with Google
- August 2010 – New article creation stopped.
- August – September 2010 – Quality review (barely 50% are rated 50 % in quality)

Tamil Wikipedia – Google Translation Project

Time line:

- September 24, 2010 – Third meeting with Google
- Oct - Dec 2010 - Translator + article selection for II phase
- Dec 2010 - II phase of project (limited articles only in userspace)
- Feb 2010 – II phase articles delivered. Review begins
- March 2011 – Google drops project (unofficial confirmation)
- June 2011 – Google drops project (indirect official confirmation)

Tamil Wikipedia – Google Translation Project

Abbreviated Time line:

- June 2009 –Dec 2009 → Google acts alone
- Dec 2009 – Aug 2010 → Phase 1 – Community involvement
- Sep 2010 – Feb 2011 → Phase 2 – Reduced project
- Mar 2011 → Google vanishes, never to be heard from again.
- Since then → Cleaning up the mess

Tamil Wikipedia – Google Translation Project

Lessons Learnt: Community involvement

- An absolute must from the beginning

- RFC's / Discussions from the start to establish guidelines/consensus

- All discussions must be on wiki and open – All goals / targets must get community's approval

- Community feedback should be considered at every phase

Tamil Wikipedia – Google Translation Project

Lessons Learnt: Become Wikipedians

- Developers / Translators must become Wikipedians
- Edit and learn; without editing you wont “get wikipedia”
- No opaque / off wiki decision making process
- Learn Wikipedia rules. Respect them.
- Talk to others on wiki

Tamil Wikipedia – Google Translation Project

Lessons Learnt: Article selection

- Community should be involved in article selection
- Source article version should be stable and of good quality – GIGO.
- Do not overwrite existing articles without consideration

Tamil Wikipedia – Google Translation Project

Lessons Learnt: Translation Quality

- Wikipedia is not a live test site for developing a translation tool

- No experimenting in Mainspace. Use Userspace

- Follow community guidelines for technical terms / transliteration

Tamil Wikipedia – Google Translation Project

Lessons Learnt: Community Again

- Feedback should be constant and continuous
- No Fire and Forget tactics. Articles should be improved continuously
- The tool is for Wikipedia and not the other way around

Tamil Wikipedia – Google Translation Project

Lessons Learnt: For Wikipedians

- Treat machine translated articles same as any other editor's
- AGF is not a suicide pact – Beware of strangers bearing gifts
- Build the community and numbers will follow