Workflow Models for Collaborative Terminology Work

Igor.Kudashev@helsinki.fi

Nordterm 2011, Vaasa, 9.6.2011
Call for collaborativeness

- Tools for collaborative content creation have rapidly gained popularity.

- In terminology work, there has been increasing pressure to pass from committee-driven work to large-scale discussions with different stakeholders.

- A good example is *Groups for IT terminology* (in Finnish *Tietotekniikan termitalkoot*).
Benefits of collaborative terminology work

- Continuous work keeps content up to date.
- More objective picture of the domain concept system and terminology due to the involvement of various stakeholders.
- A chance to preserve and activate "silent knowledge" and "grass level knowledge" in organizations and companies.
- Higher speed of content creation.
- Lower direct costs of content creation in respect of the content volume.
Overview of the *TermFactory* project

- A part of a three-year project *ContentFactory* carried out at several departments of the University of Helsinki and Aalto University.

- Financed by Tekes and a number of language industry companies.

- Aimed at creating architecture and a workflow model for collaborative ontology-based terminology work.

- One of the outcomes is *Quality Manual for Collaborative Terminology Work*. 
The initial "terminologist-oriented" model

- Stage 1: general users comment using a forum.
- Stage 2: moderators review comments and modify records on a Wiki.
- Stage 3: terminologists review the Wiki and modify OWL repositories with an ontology editor.
Disadvantages of the terminologist-oriented model

- Terminologist-oriented model is an example of *supervisor moderation* typical for Internet forums.

- Disadvantages from the point of view of collaborative terminology work:
  - Too slow, as new term candidates and modifications to term records have to be checked and approved by at least two people.
  - The original proposal by a general user may get distorted as it passes two stages of interpretation and revision.
  - An “authoritarian” model, as moderators enjoy trust by the administrator but not necessarily by the community.
Principles of a “community-oriented” model

- Direct and real-time editing of term records is available to all the members of the community.

- Trust by administrators is counter-balanced with trust by the community.

- Quality is achieved with the help of user moderation, reputation system, and content rating.
Strict authentication and other security procedures

- Registration under one’s real name, persistent login.
- Verification of registration by SMS for individual users.
- Batch registration by an organization’s system administrator for corporate users and a “verified corporate user” badge for them.
- Users have to provide detailed information about their expertise in terminology work, multilingual communication and special domain expertise.
- Users’ IP-addresses are registered.
Supervisor moderation preserved only as a means against behavioral problems

- Moderators address behavioral problems, such as insulting posts, personal attacks, spam and other kinds of abuses.

- Contributors can report using a *Signal to moderator* button.

- Administrator appoints moderators based on users’ requests for moderatorship.

- Wikipedia has elaborated a good set of criteria for the selection of moderators.
Direct and real-time editing

- Direct and real time editing of term records is available to all the registered members of the community.

- Previous versions of records are stored in *Edit History* and can be reverted from there (cf. Wikipedia).

- No entry locking but a notice is issued if a record is being edited by another user.
The purpose is to estimate the results of collaborative terminology work by ranging different versions of the same field.

Contributors can rate content by assigning one to ten stars to a data field in the current version of a record or its previous versions in the edit history.
Content rating 2/2

- Users can rate an unlimited number of fields.

- They can change their rating for a particular version of a field but they cannot rate it multiple times.

- Users cannot rate their own contributions.

- Ratings of the data fields can be made visible to the users upon request.

- It is desirable that the system should be able to assemble term records from the versions of the fields with the highest rating.
Reputation system

The reputation system was inspired by *Slashdot*
- A news and commentary site dedicated to technology issues, especially open source software.
- One of the oldest and most popular collaborative weblogs in existence.
- Represents the definitive example of a user moderation system in the context of Internet forums (Wikipedia).
- A highly sophisticated reputation system has been tested and fine-tuned for a long time.

The *Slashdot* reputation system was adapted and modified for the needs of collaborative terminology work.
Trust by community 1/4

- Users rate each other’s contributions and in this way gain positive or negative “karma”.

- Each contributor gets a certain amount of karma points on a regular basis.
  - At higher reputation levels users get more karma points to distribute.
  - Karma points do not expire but new points are not distributed until the old ones have been used.

- Users give their karma points to other contributors for the contributions they have particularly liked or disliked.

- Contributors are encouraged to focus more on rating good work (70 % of points have to be positive).
Trust by community 2/4

- Karma points cannot be assigned to contributors directly but only by rating their contributions.

- When selecting the number of stars for a field, users can also assign a positive or negative karma point to the person who has last contributed to the field in question.

- Contributions can be rated both in the actual version of the entry and in the edit history.

- Users can rate the same contribution only once and only with one point.

- Karma points are anonymous.
Trust by community 3/4

- Karma points are domain- and language-specific.
  - Each term is linked to some domain, and each data field is linked to some language section, so the system can automatically keep track of how many points the user has got in each domain and language.

- This information is later used for determining in which domains and languages the user can become a verifier.

- Each karma point assigned to the user is divided into two halves – domain-specific and language-specific karma.
Having gained enough karma points, contributors are promoted to the next reputation level.

Each new level increases contributors’ authority as they are awarded with more karma points to distribute.

At the same time, the higher the reputation level, the more points are required for an upgrade to the next level.

Having reached the highest reputation level (e.g. 10), contributors get the status of distinguished contributor.

This allows them to verify term articles in domains and languages in which they have gained enough karma points.
Trust by administrators

- Professional terminologists and other experienced contributors can be verified by the administrator and get the status of **authorized contributors**.

- Authorized contributors can verify term records according to their declared language and domain proficiency.

- Authorized contributors and distinguished contributors have badges of different colours.
Demotion

- Contributors who get negative karma can be demoted to lower reputation levels.

- In case of demotion from the top level, contributors lose their right to verify data.

- If they are unable to restore their reputation level within a month, verifications made by them are deleted from the system.
Data verification 1/2

- Distinguished and authorized contributors can verify individual data fields or groups of data fields in a term record.

- “Standard verification”: verifier has used the best practices of systematic terminology work (for example, concept analysis) when verifying term records.

- “Quick verification” is used in cases when a limited amount of terms which do not necessarily form a concept field have to be quickly verified by an experienced member of the community.
Data verification 2/2

- Verified data is not copied to a separate collection; it is only marked as verified.

- Even if a verified version is later overwritten by other contributions, it can still be found in the edit history.

- At least the following data should be provided in the “verification ticket”:
  - type of verification (quick or standard), verifier, organization which the verifier possibly represents, verification date and verifier’s comments.

- For a particular field, only the latest verification ticket is stored while previous verification tickets of the same verifier are deleted.
Closed collections

Closed collections are open to the administrator of the collection and users invited by him (her).

If administrator of a closed collection invites other users, s/he can no longer terminate the collection without their consent.

Records can be created from scratch or copied from the public collection.

Closed collections can be made public by the administrator of the collection with the consent of the community.
Reputation system in closed collections

- The same as in public collection but
  - Administrator may choose not to use the “karma model” but to manually assign the status of authorized contributor to selected users.
  - Administrator can customize the amount of karma points for the purposes of the closed collection.

- Karma points gained in closed collections do not count in the public collection.
Workflow model for traditional terminology work

- Always done in closed collections.

- Users work on records in turn, according to the plan; the entry is locked from others.

- No reputation system; user roles (such as "terminologist", "domain expert", "editor", "proofreader") are assigned by the administrator.

- Content rating may be useful.
Future prospects

- The work continues in the CF-PreComm project (2011-2012).
- Proposed model should be tested, evaluated and fine-tuned in real terminological projects.
- A wider range of workflow models and "ecosystems" for different clients and projects.
  - E.g. small terminology project in a trustworthy environment vs. wide-scale collaborative project; descriptive vs. normative terminology project.
- *Workflow Guide* to help users choose a workflow that suits their needs the best.
Thank you for your attention!

Igor.Kudashev@helsinki.fi