TermFactory: A Platform for Collaborative Ontology-based Terminology Work

Igor Kudashev, Irina Kudasheva and Lauri Carlson
University of Helsinki

TermFactory is an array of standards and tools based on Semantic Web ontology techniques. Its mission is to allow companies, organizations and individual contributors to collaboratively produce multi-domain special language vocabularies and ontologies. Ontologization of terminological data has several benefits, such as global identification of concepts, automatic checks for logical errors, reasoning and data propagation, presentation of data in machine readable and -processable form and the possibility to substitute static entries with dynamic 'views' tailored according to the user's needs and preferences. Collaborative work is a double-edged sword which potentially has many benefits but may also present serious challenges. In our poster, we describe challenges of collaborative terminology work and possible solutions to them. If well-organized, a collaborative project can be quite successful, as the example of Wikipedia and many other collaborative projects on the Internet demonstrate.

1. About the project

TermFactory (TF) is a part of a larger project ContentFactory (2008–2010) carried out at several departments of the University of Helsinki and Helsinki University of Technology. The project is financed by the Finnish Funding Agency for Technologies and Innovation (Tekes) and a number of language industry enterprises. The TermFactory workpackage of the project aims at creating a platform and a workflow for distributed collaborative ontology-based terminology work.

2. TermFactory's mission

TermFactory is an array of standards and tools based on Semantic Web ontology techniques. Its mission is to allow companies, organizations and individual contributors to collaboratively produce multi-domain special language vocabularies and ontologies (see *Figure 1*).

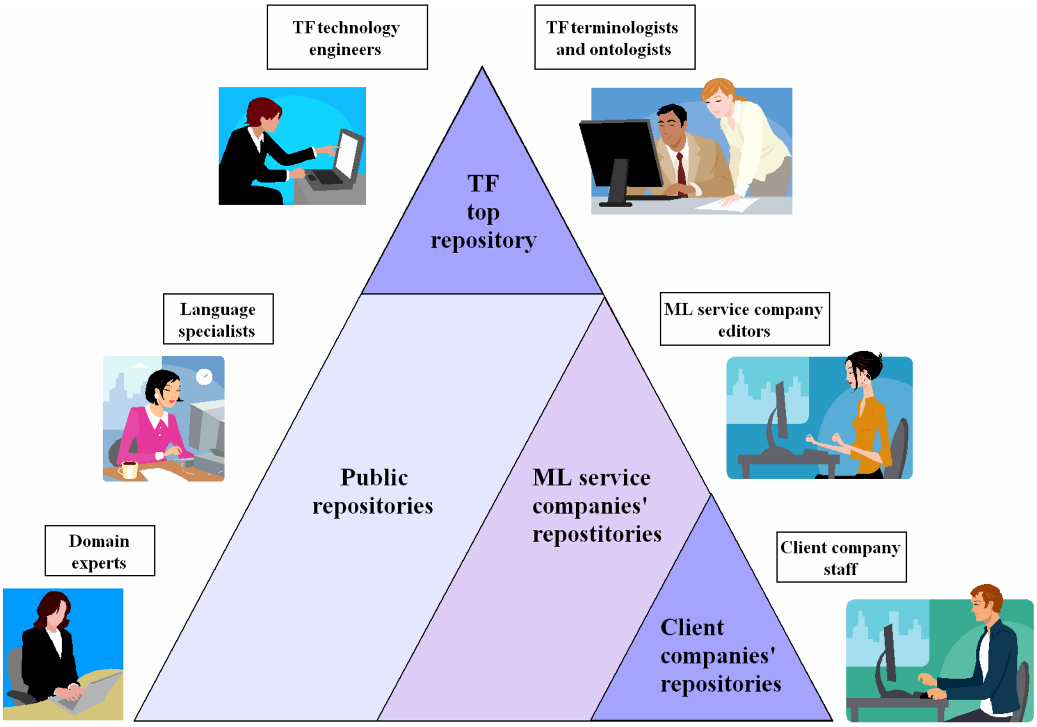
TF technology engineers
TF terminologists and ontologists
TF top repository
Language specialists
ML service company editors
Public repositories
ML service companies' repostitories
Domain experts
Client company staff
Client companies' repositories

Figure 1. TermFactory's layers and users

TermFactory can be used to organize content and standardize communication in global multilingual organisations as well as to boost exchange of ideas and innovations and support education across language barriers.

3. TermFactory architecture

TermFactory is designed as a distributed resource. TermFactory network consists of OWL repository servers and collaborative wiki / forum platforms connected by a common directory (see *Figure 2*). The nodes communicate in a peer-to-peer fashion on the web service layer. Collaborative platform servers are loosely coupled to the TermFactory repositories.

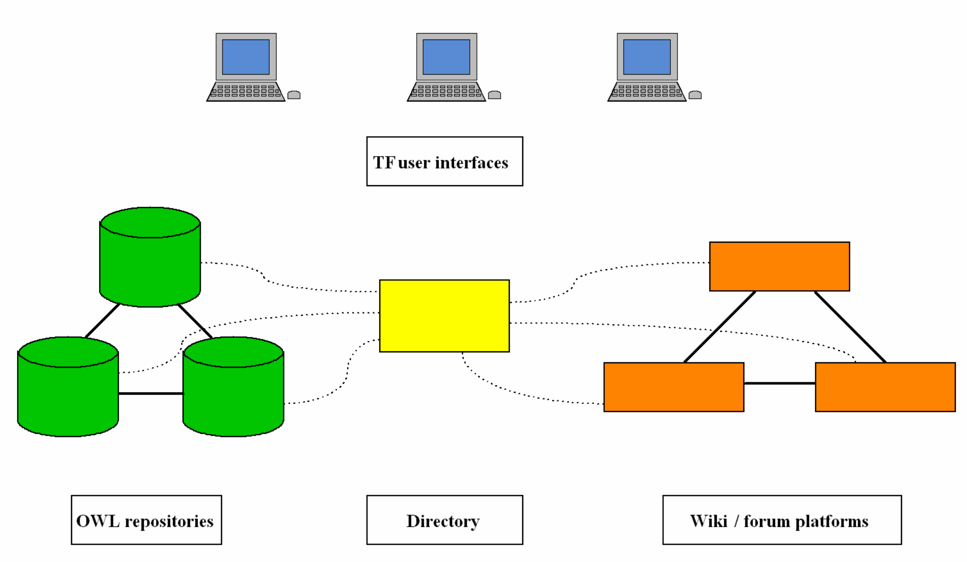

Figure 2. TermFactory architecture

4. Semantic web technologies

Ontologies and other Semantic Web technologies lie at the very core of the TF platform. Terminological resources are stored in TermFactory in the form of OWL ontologies either as web documents or database repositories.

Ontologies are formal, machine-readable and -processable descriptions of classes of objects and their relations in a domain of discourse. Ontologization of terminological data has several benefits.

Ontologies can be a powerful instrument of creating and sharing common understanding about concepts. Defining a concept requires some thinking and negotiations between parties who are going to commonly use it but thanks to this process the parties can in the future be much more certain that they are referring to the same, globally identified concept.

Ontologies can be automatically checked for logical errors. This helps finding mistakes and inconsistencies in the existing and newly created terminology. When information contained in multilingual terminological resources is converted into an ontology and validated both by a reasoner and a team of experienced terminologists and ontologists, it can be used by machines to better ‘understand’ the world and natural language expressions.

With the help of reasoning computers can also infer knowledge which is expressed only implicitly. For example, as antonymous relations are symmetric, there is no need to add antonyms crosswise in two entries as the second operation can be done automatically by the system. For more loose types of relations which can not be inferred directly the system can make assumptions which need to be verified by a human user. Both types of propagation speed up input of data in a lexicographical processor and assist automatic management of links.

Another lexicographically interesting opportunity provided by ontologies is that they allow to get rid of the notion of an entry as the mandatory rigid 'container' that keeps individual data elements together. Indeed, if each object of description and each element of the description are represented as individual objects, whose relations with other objects are explicitly and unambiguously described, the same data can be presented to the end users in many different ways. This means that static entries can be substituted by dynamic 'views'.

5. Collaborativeness

TermFactory is designed to support collaborative, Wikipedia-like terminology work by communities. The tools for collaborative terminology work are based on the content management systems like Mediawiki. They can also be implemented as plug-ins for the wiki platforms already used by companies, organisations or web communities.

Collaborativeness is a broad concept. For the purposes of the TermFactory project the following definition of collaborative terminology work has been adopted:

> *Collaborative terminology work is terminology work conducted in an electronic environment by two or more users in such a way that users can edit directly and without prior notice contributions of other users*.

6. Benefits and challenges of collaborative approach

Well-organized collaborative work has many benefits, the most important of which are the following:

- High speed of content creation.
- Continuous character of the work allows keeping content up-to-date.
- Use of electronic environment allows users to communicate across space and time.
- Large-scale cooperation, common 'wisdom of crowds' may result in good or even excellent quality.
- 'Silent knowledge' and 'grass level knowledge' get preserved in organizations and companies.
- Lower costs in respect of the content volume.

At the same time, collaborativeness presents many challenges which have to be addressed when designing a platform for collaborative terminology and ontology work. Below are discussed possible solutions to the main challenges of collaborative terminology work.

Problem: Lack of skills and experience in terminology work and lexicography among the participants.

Possible solutions:

- Clear instructions and help.
- Best practice guidelines.
- Tutorials.
- Supervision.
- Division of labour.

Problem: No general model for creation of terminological collections can be provided due to different backgrounds and needs of the participants. This may result in very heterogeneous content and complicate the organization of search.

Possible solutions:

- Elaboration of minimum requirements to terminological description (mandatory fields).
- System of reminders for filling in important fields.
- Templates and models of terminological articles and collections, a plan for creation a terminological collection, source documentation, etc.
- Top-level hierarchy of data categories.
- Core domain classification.
- Common register of language and country codes.

Problem: A great, potentially unlimited number of languages; no clear source and target language(s); combining several units of description which represent slightly different concepts may result in overloaded and messy entries.

Possible solutions:

- It should be possible to customize views in such a way that only those languages and types of information are made visible in which the user is currently interested in.
- The object of description should be terminological lexeme (a set of forms of an LSP unit sharing a common meaning).
- In the ‘edit view’ only information related to the described unit should be added.
- The default ‘browse view’ should also consist of the described terminological lexeme and its description. Full description of related units (synonyms, equivalents, etc.) can be opened in a new window/tab/pane, etc. Partial description without cross-language equivalents can be opened as an insert in the current view.
- Other types of views (concept-oriented, term-oriented, etc.) should be available on demand.

Problem: There is no person who makes the final decision about the content. This can lead to ‘edit wars’.

Possible solutions:

- Request for comments/votes from other users to help resolve the dispute if an edit war has started.
- ‘Call moderator’ button on the discussion page of each term article.
- Limitation of reverts for a given period of time.

Problem: Users’ identity is usually not verifiable which may lead to identity frauds and vandalism.

Possible solutions:

- Compulsory registration under one's real name.
- Verification of enterprise users by companies.
- Verification of individual users' registration by SMS (account activation code is sent to mobile phone).
- Users' IP-addresses should be registered in the system.
- Automatic alerts to moderators if most of the text in a term article has been removed.

Problem: Freedom of speech can cause controversy between users with different backgrounds.

Possible solutions:

- All significant views should be represented fairly, proportionately, and without bias (the so-called *'neutral point of view'* principle in Wikipedia).

Problem: Users may unintentionally or intentionally publish materials violating somebody's copyright and/or materials that do not comply with national laws.

Possible solutions:

- Upon registration new users should accept the terms of the licence agreement with providers of the service stipulating among other things that it is users' responsibility to make sure they do not violate other authors' copyright and do not breach national laws by publishing material in the system.
- The importance of copyright and legal compliance of the published data should be stressed in the rules, tutorials, etc.
- Reporting copyright abuse to moderators should be easy.
- Users should be obliged to carefully document sources and borrow text in such a volume that it corresponds to citation.
- Indication of sources should be made easy for the users.

Problem: Copyright to terminological content created collaboratively.

Possible solutions:

- To prevent disputes over authorship users should not have copyright on collaboratively created content.
- Commercial use of collaboratively created content in public collections is not allowed.

Problem: Lack of motivation to use collaborative terminology work in organizations and companies.

Possible solutions:

- Internal marketing of the benefits of collaborative terminology work.
- Low barrier to start using the system: ready solutions and templates for novice users, intuitive and user-friendly interface, etc.
- Reputation system with visible prestige levels and more authority on higher levels (e.g. more points to rate others' work, the right to verify data).
- A system of awards (e.g. badges for high quality contributions).
- Occasional extra rating points to stimulate the activities.
- A salary raise for active contributors.

7. Conclusion

TermFactory is an array of standards and tools based on Semantic Web ontology techniques. Its mission is to allow companies, organizations and individual contributors to collaboratively produce multi-domain special language vocabularies and ontologies.

Ontologization of terminological data has several benefits, such as global identification of concepts, automatic checks for logical errors, reasoning and data propagation, presentation of data in machine readable and -processable form and the possibility to substitute static entries with dynamic ‘views’ tailored according to the user’s needs and preferences.

Collaborative work is a double-edged sword which potentially has many benefits but may also present serious challenges. If well-organized, a collaborative project can be quite successful, as the example of Wikipedia and many other collaborative projects on the Internet demonstrate.