



DELIVERABLE

Project Acronym: thinkMOTION

Grant Agreement number: 250485

Project Title: Digital Mechanism and Gear Library goes Europeana

D8.1: Translation workflow is improved

Revision: 5

Authors:

Jean-Christophe Fauroux (IFMA)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
1	20/06/11	J.C. Fauroux	IFMA	Introduction, objectives, general aspects
2	23/06/11	U. Döring	IUT	Description of examples
3	24/06/11	J.C. Fauroux	IFMA	Selected improvements, further work
4	15/07/11	U. Döring	IUT	More detailed examples and snapshots
5	18/07/11	R.Brecht	IUT	Review

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Content

1	Introduction	4
2	Objectives	4
3	General aspects of the improvement of the translation workflow	4
4	A more detailed view onto selected improvements	6
5	Further work planned in WP 8 in the next year	10
6	References	11

1 Introduction

The thinkMOTION project [1] has the ambition to make the DMG-Lib portal [2] to the major digital library of machines and mechanisms at the world level. As an international database containing a lot of textual content, thinkMOTION integrates a work package in charge of multilingual aspects. Although most researchers speak scientific English, it is important that the content remains perfectly understandable to a large audience, particularly the non-specialists, also concerned by this project.

2 Objectives

According to the abilities of the project partners the targeted languages were defined to be English, German, French, Italian, Spanish and Romanian. Objectives are at two levels:

- translating the portal and where needed the internal pages
- translating contents to allow multilingual searching and browsing

Four tasks were defined in the work package WP8:

- 8.1 Improvement of translation workflow
- 8.2 Translation of web pages
- 8.3 Translation of metadata
- 8.4 Maintenance of a thesaurus

Deliverable D8.1 (translation workflow is improved) which is described in this document is the direct result of task 8.1.

3 General aspects of the improvement of the translation workflow

The thinkMOTION project is settled on the existing software and data basis of the DMG-Lib Project, a research project initiated in Germany seven years ago. Initially only in German and partly in English, DMG-Lib is in the thinkMOTION project enlarged to a wider range of languages to enhance its attractiveness by covering more languages spoken by the citizens of European countries.

Task 8.1 (resulting in deliverable 8.1) has the ambition to define a new workflow capable to generate international content, using the capacities of all the partners in the project as effective as possible. The task has two main aspects. The first covers the general design, the discussion as well as the practical evaluation of the improved translation workflow whereas the second aspect is related to the implementation of the according software tools and their integration into the production database.

The design of the improved workflow as well as some new software tools were presented in September 2010 during the kick-off in Ilmenau (Figure 1). The overall structure of the workflow shown in this figure can be divided into two columns. In the left part the required workflow to translate the metadata attached to items is shown. These metadata can be divided into two categories:

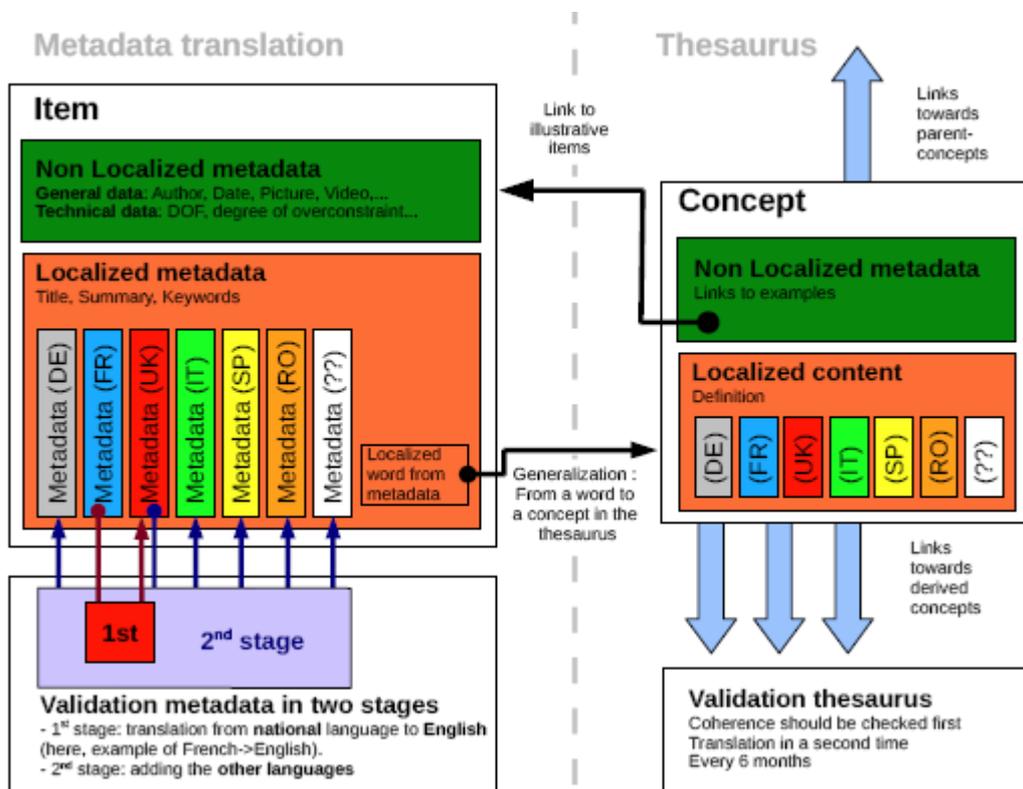
- Non localized metadata (on the top): those are not expressed in a language and consequently will not require any translation. In a scientific project such as thinkMOTION, a lot of metadata is non localized, e. g. numerical data describing the considered machines or mechanisms (like the number of bodies, reduction ratio, etc.). The names of inventors, authors, places etc. will also usually be kept unchanged. Finally, multimedia data such as pictures or videos do not have to be translated either, provided they do not include text writings.

- Localized data (on the bottom) such as the title or the text of a summary are worthwhile to be translated. For the moment, 6 languages are considered but the provided architecture is easily extendable to further languages.

The translation process was thoroughly separated into two stages:

- First stage: priority is given to translation from the mother tongue towards English, which is the most widely used language in science. The quality of translation at this stage is critical, as the results will be re-used later for subsequent translations. There is a strong validation stage here at the end of the translation.
- Second stage: this consists in translating from English to other remaining languages, different from the initial mother tongue of the resource. This stage may become a heavy task (many languages in parallel). It was decided to give it a lower priority.

Figure 1. Translation process in the thinkMOTION project.

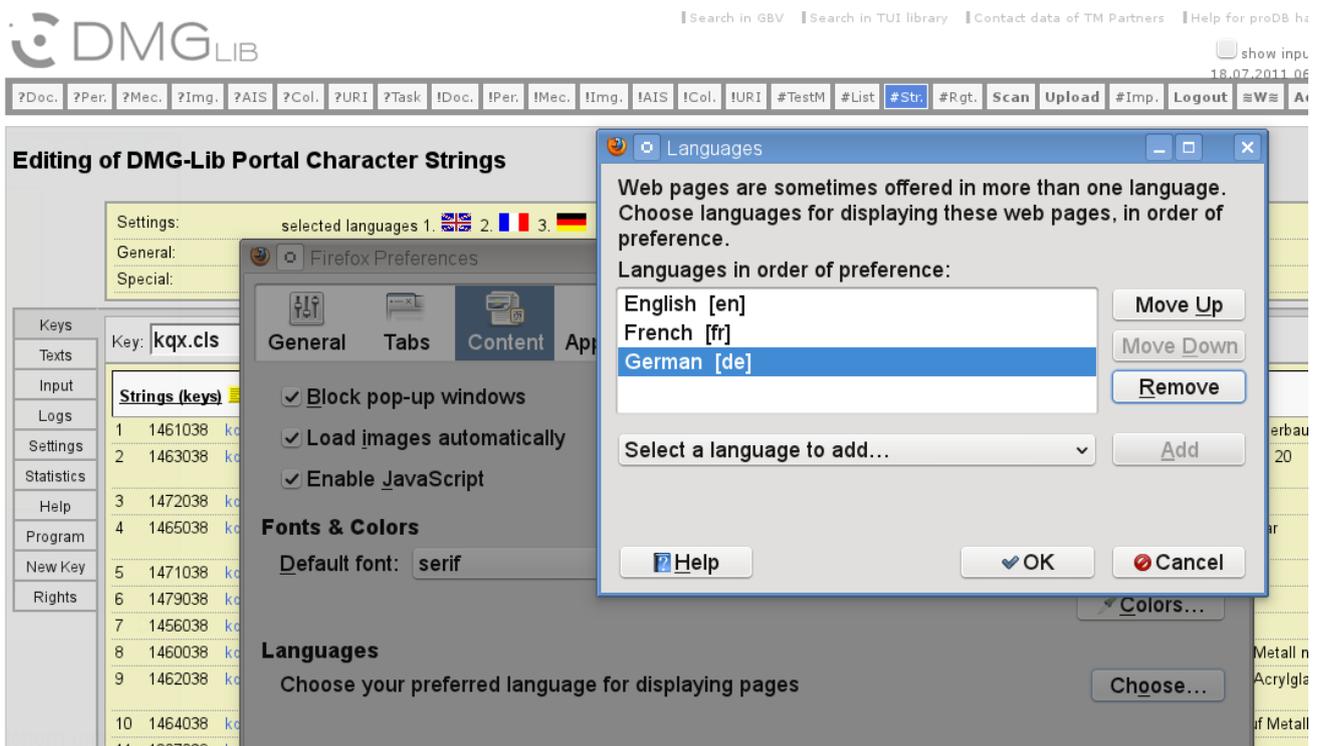


In the right part of Figure 1 the thesaurus related aspects are shown. The ambition of thinkMOTION is to provide a strong support for a deeper comprehension of all the main concepts involved in machine engineering. A thesaurus will be built, based on acknowledged resources such as the IFToMM terminology [3]. The same translation process in two stages will occur here. Moreover, the thesaurus is represented as a semantic network and hypertext navigation is expected in it. This requires a hyper-graph structure between concepts but also edges connecting concepts to items, and this in several languages. Current work handles the computer representation of the thesaurus.

4 A more detailed view onto selected improvements

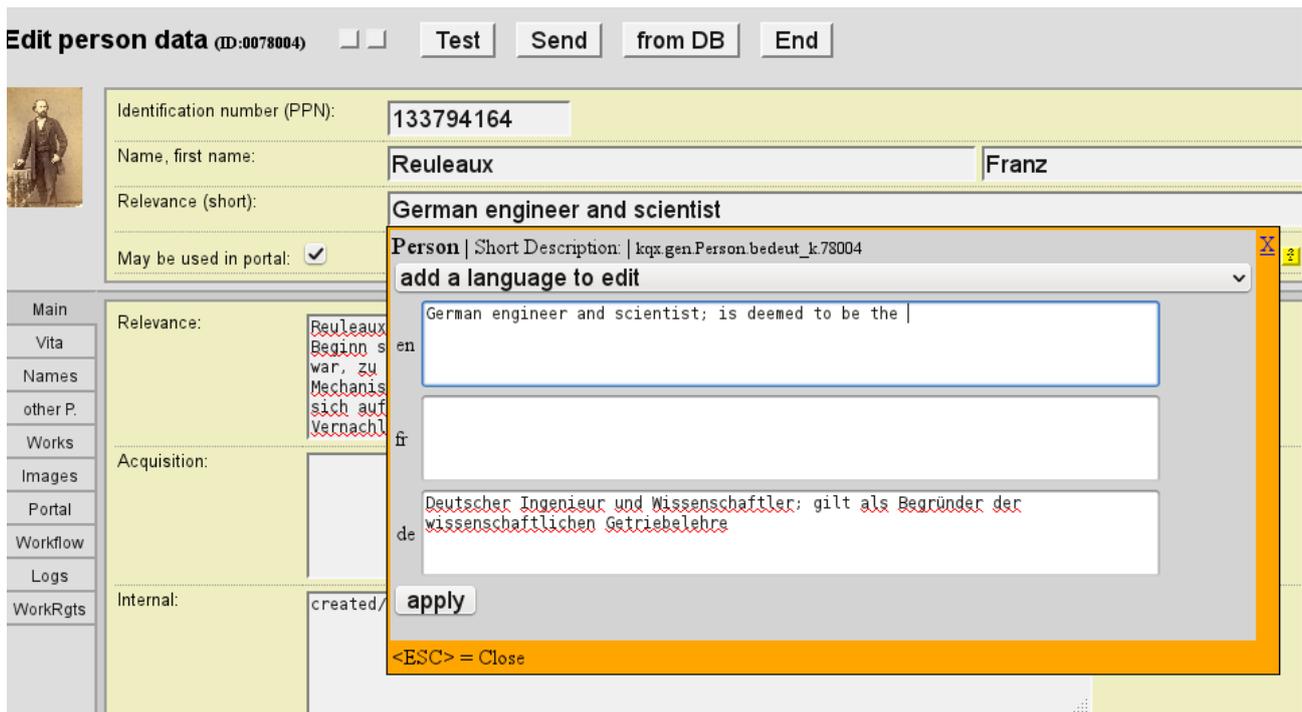
First of all an improved version of detecting user preferences from browser settings was introduced. It is based on the language settings in the browsers which allow each user to specify an ordered (or even weighted) list of languages he prefers when reading page content. See Figure 2 which shows such a selection in Firefox. In the figure English is preferred, but French as well as German will also be accepted by the user. The information is extracted from the HTTP-headers and used to preselect the set of languages the translating users can edit in the translation windows (Figure 3).

Figure 2. Selecting certain languages in the browser (here Firefox)



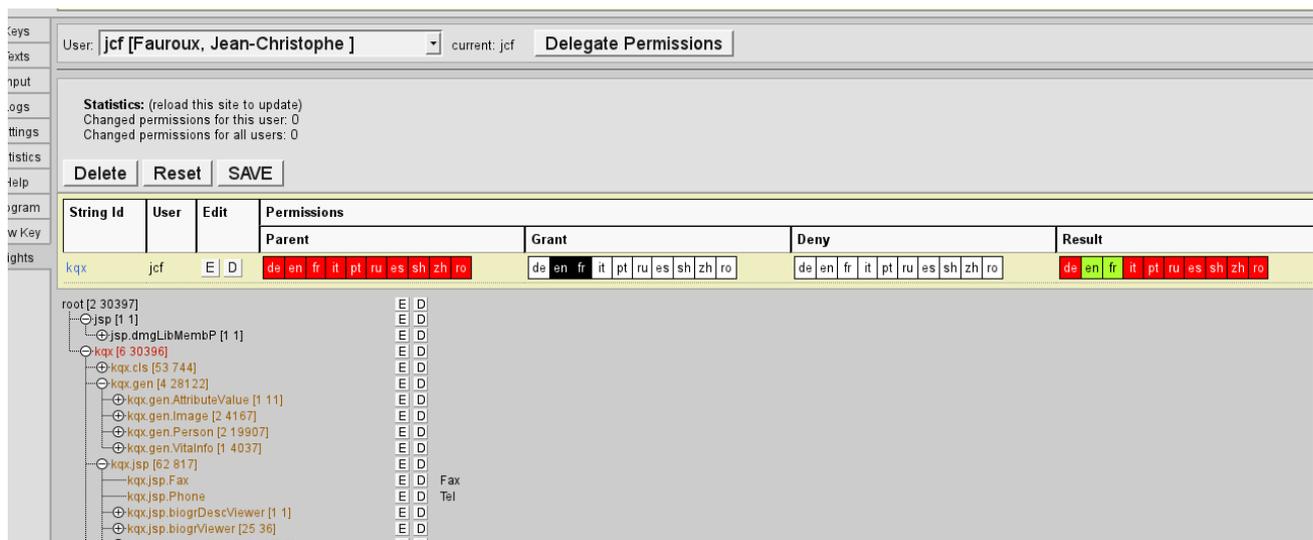
The translation window for textual content opens when an according input field is clicked. Figure 3 is a snapshot which shows the translation window after clicking on the text field for the “short relevance” of the person Franz Reuleaux (the current content in this field “German engineer and scientist” is filled with the best matching languages – in this example English). The editor is opened providing input fields in the 3 selected languages (boxes for the languages English, French and German as shown in Figure 3). Nevertheless the editing staff member may add further language boxes. As far as he has the permission he can also edit those languages, else the texts are read-only for him. After sending the changed dataset the entered data is indexed in the search machine and can be seen in the portal.

Figure 3. Extending metadata for Reuleaux by completing the English short description.



To handle the access to potentially hundreds of thousands of texts with different languages an improved permission management was introduced. The new data structures allow to order texts (e.g. words, phrases or larger text blocks) as text nodes in a text tree. The internal name of each node is used to determine its position in the tree (Figure 4).

Figure 4. Defining permissions for a user



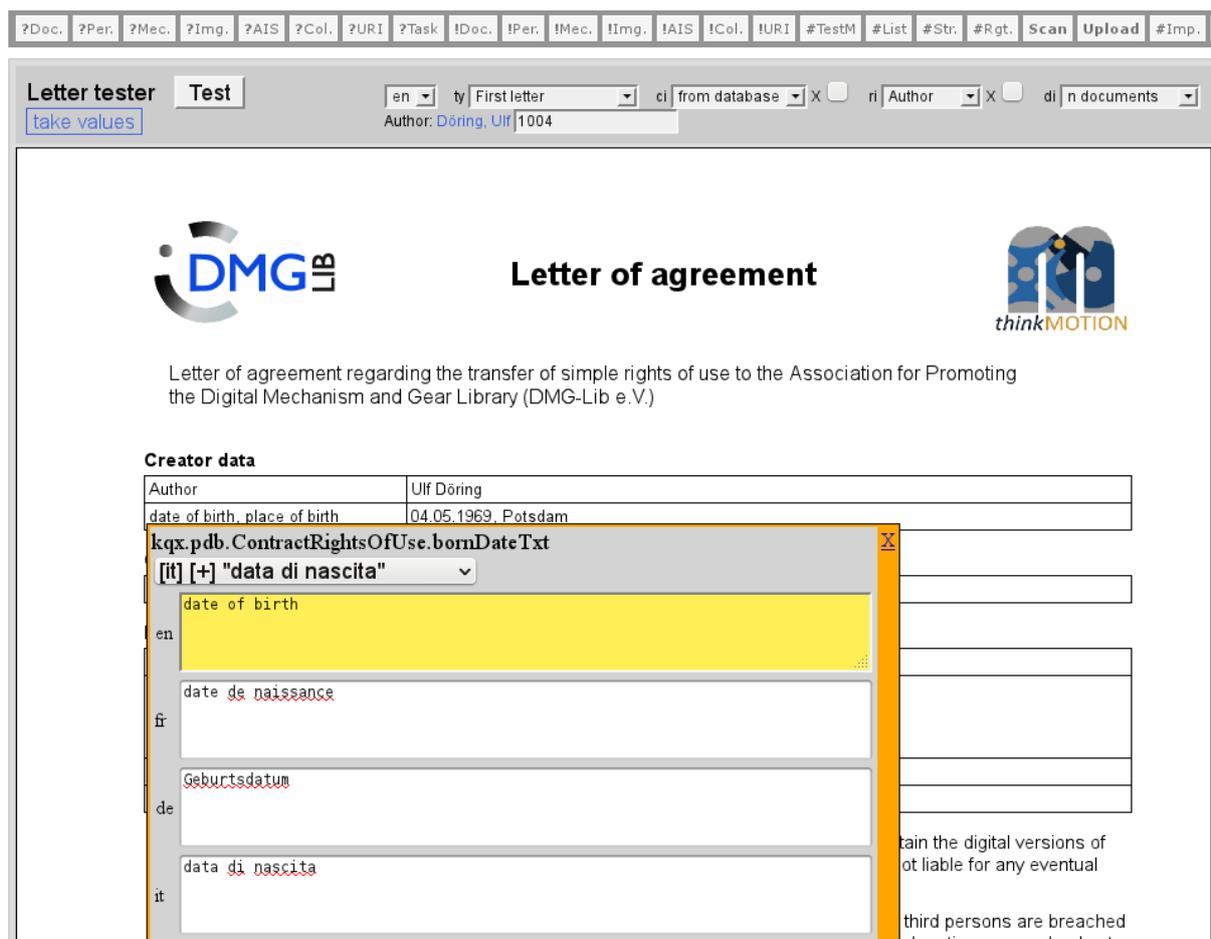
For each node in the tree the edit permission can be granted and denied according to certain users and languages. Based on the hierarchical structure permissions are passed top-down. I. e. when a

permission definition is to be retrieved for a certain node N then we have to use the permission defined in node M which is nearest to N going up starting at N (of course M may be equal to N). In Figure 5 the permissions of the user “jcf” are shown which has the permission to edit all English and French texts in nodes under the node with the name “kqx.”.

A tool helps to administrate these permissions (e. g. copy permissions of one user to another). Because groups of users have often the same or at least similar permissions this tool supports groups too and makes this way the management work more effective. A further feature of the administration tool shown in Figure 4 is the statistics which calculates e. g. the number of sub-nodes and allows to estimate the necessary effort/progress of translation for this node (and the sub nodes).

Figure 3 shows the newly introduced editor for item metadata. In Figure 5 the editing of quasi static texts is shown. Those texts do not belong to certain items. In the shown example a text used in the letter of agreement is edited. The edit window was opened by a click on the text “date of birth” in the table with the creator data. As in the other examples the English, French and German texts are shown by default, but here the user added Italian too.

Figure 5. Defining texts for the letter of agreement.



The letters of agreement in their different versions are a good example for very project-critical texts. Those letters will be signed by rights holders (authors, heirs, institutions etc.) and the meaning of the

translated text blocks must not differ. Else automatic rights handling would fail and the resulting erroneous rights metadata or visibility of items could cause lawsuits.

The letters of agreement are also a good example for the message system according to needed translations. When for example a text is not defined in the target language (e. g. Italian for an Italian author) then a message is shown that the letter can not be generated and the user (the staff member) is directed to this editor.

Within the translation workflow different paths can be used to translate texts. Two main paths can be distinguished. One is based on the export and import of CSV-files (files with comma-separated values). This path is used mostly at the begin of the translation for a certain set of web pages (more general: a set of texts belonging together according to some criteria). Later masks as shown in Figure 7 are used or more special translations (e. g. as already shown in Figure 5).

The lists of texts (CSV-files) are exported and imported via JSPs (Java Server Pages) in the internal production database. Those CSV-files are usually not processed directly by the translators. A conversion to/from a sheet format like ODS (OpenOffice Calc) or XLS (Microsoft Excel) is advantageous.

Figure 6. Import form for the control of mass imports

111 kqx cls modelInputMovementSemantics.l.undefvalue	<input type="checkbox"/> unbestimmter Wert	<input type="checkbox"/> undefined value	<input type="checkbox"/> valeur non définie
112 kqx cls modelInputMovementSemantics.s.frame_motion	<input type="checkbox"/> Führen	<input type="checkbox"/> Frame motion	<input type="checkbox"/> Pilotage
113 kqx cls modelInputMovementSemantics.s.multiple	<input type="checkbox"/> Mehrere Arten	<input type="checkbox"/> Multiple	<input type="checkbox"/> Multiples
114 kqx cls modelInputMovementSemantics.s.orientating	<input type="checkbox"/> Orientieren	<input type="checkbox"/> Orientate	<input type="checkbox"/> Orienter
115 kqx cls modelInputMovementSemantics.s.positioning	<input type="checkbox"/> Positionieren	<input type="checkbox"/> Position	<input type="checkbox"/> positionner
116 kqx cls modelInputMovementSemantics.s.rest	<input type="checkbox"/> sonstige	<input type="checkbox"/> other	<input type="checkbox"/> autre
117 kqx cls modelInputMovementSemantics.s.rot_trans	<input type="checkbox"/> Dreh-Schieben	<input type="checkbox"/> Rotation superposed by rectilinear translation	<input type="checkbox"/> rotation combinée avec une translation rectiligne
118 kqx cls modelInputMovementSemantics.s.rotation	<input type="checkbox"/> Drehen	<input type="checkbox"/> Rotation	<input type="checkbox"/> Rotation
119 kqx cls modelInputMovementSemantics.s.screwing	<input type="checkbox"/> Schrauben	<input type="checkbox"/> Screw motion	<input type="checkbox"/> Vissage
120 kqx cls modelInputMovementSemantics.s.translation	<input type="checkbox"/> Schieben	<input type="checkbox"/> Rectilinear translation	<input type="checkbox"/> translation rectiligne
121 kqx cls modelInputRevolutionSemantics.l.not_revolute	<input type="checkbox"/> nicht unlaufender Antrieb	<input type="checkbox"/> new: not rotating drive <input checked="" type="checkbox"/> old: not continuously rotating drive	<input type="checkbox"/> entrainement non rotatif
122 kqx cls modelInputRevolutionSemantics.l.rest	<input type="checkbox"/> Sonstiges	<input type="checkbox"/> Others	<input type="checkbox"/> autres

A snapshot of a part of an imported table is shown in Figure 6. Here (based on the German and English texts) French texts were added and some English and German texts corrected. The state shown in the figure is between a first import and a further corrective import. Usually the complete light yellow lines would be hidden (because they are completely equal to the database entries), but here we ticked a check box in the HTML-form to give an example for an imported list of texts. The (orange) highlighted field marks a difference which should be checked. By default the “do import this text”-flag is ticked for all the changed texts because we can assume that most changes will be

correct. Thus, if every change is OK then the importing person has only to press the “send”-button and the import into the database is performed by the software.

There are different parameters which can be used to control the import, e. g.:

- the set of imported languages,
- the number of rows to be skipped (e. g. a row for the serial number),
- the number of lines to be skipped (e. g. comment lines),
- the used separator characters,
- the user who has generated the imported file (this user is often different from the importing person) etc.

The import table is a further example for consistency checks. Here the staff member which controls the import can visually compare the length of the texts to detect empty or incomplete translations or even a wrong use of parameters.

Figure 7. Improved “old style editor” usable to enter or change texts.

The screenshot displays the 'Editing of DMG-Lib Portal Character Strings' interface. At the top, there are settings for selected languages (1. UK, 2. FR, 3. DE), general parameters (def:30397, txt:15830, en:3044, de:10165, fr:1019), and special parameters (ID:1205038, key:kqx.cls.DmgSet.set.basemech.wedge, str:Keilschubgetriebe). Below these are three text input fields for English, German, and French translations. The English field contains 'Wedge mechanism', the German field contains 'Keilschubgetriebe', and the French field contains '@'. A 'Send' button is located below the input fields. At the bottom, there is a log table with the following data:

ID	kqx.cls.DmgSet.set.basemech.wedge	old	new	Time	User
1	en		wedge	29.01.2008 17:26	Reeßing, Michael (reessing)
2	en	wedge	Wedge mechanism	18.02.2008 13:25	Brecht, Rike (rike)
3	de		Keilschubgetriebe	29.01.2008 17:26	Reeßing, Michael (reessing)

Figure 7 shows the current version of the editor which was initially developed to translate terms in the national DMG-Lib project (mainly from German to English). Below the 3 edit boxes a table with log entries is placed. Especially the visualization of those logs was important to be adapted to the new needs.

5 Further work planned in WP 8 in the next year

- translation of the remaining web pages (especially in Italian / Spanish / Romanian) until end of 2011

- translation of internal production database pages according to demand
- July-September 2011: joint work IFMA-IUT to check the translation on typical items. Further improvements and redaction of the tutorial for the translation workflow
- September 2011: start of mass translation for metadata (8.3)
- September-December 2011: exchanges IUT-IFMA about the software architecture for the thesaurus (based on the IFToMM dictionary as described in [4]). Providing a list of terms to the partners for translation. Integration of UPT term list into the thesaurus.

6 References

1. thinkMOTION project web site. <http://www.thinkmotion.eu>
2. DMG-Lib portal. <http://www.dmg-lib.org>
3. IFToMM Terminology. <http://www.iftomm.3me.tudelft.nl>
4. Döring U., Brix T., Corves B.: Suggestion for a More Productive Workflow and Infrastructure of the Permanent Commission on Standardization of Terminology. In: Proceedings of 23rd Working Meeting of the IFToMM Permanent Commission for Standardization of Terminology on MMS, Minsk/Gomel, Belarus, 21-26 June 2010, pp.49-57 (2010) Available via DMG-Lib. <http://www.dmg-lib.org/dmglib/handler?docum=10677009>