TmxSQL

Managing Translation Memories with SQL

Steve Braich

https://github.com/stevebpdx/TmxSql

Managing Translation Memories

Background:

- Translation memories (TMs) are saved translations that can be leveraged later for future translations.
- They are usually saved in an XML file format known as TMX (Translation Memory Exchange)
- There are virtually no open source TM desktop management tools that have continued support.
- Translation memories are a very valuable asset for training machine translation models.

Why the need for TM Management?

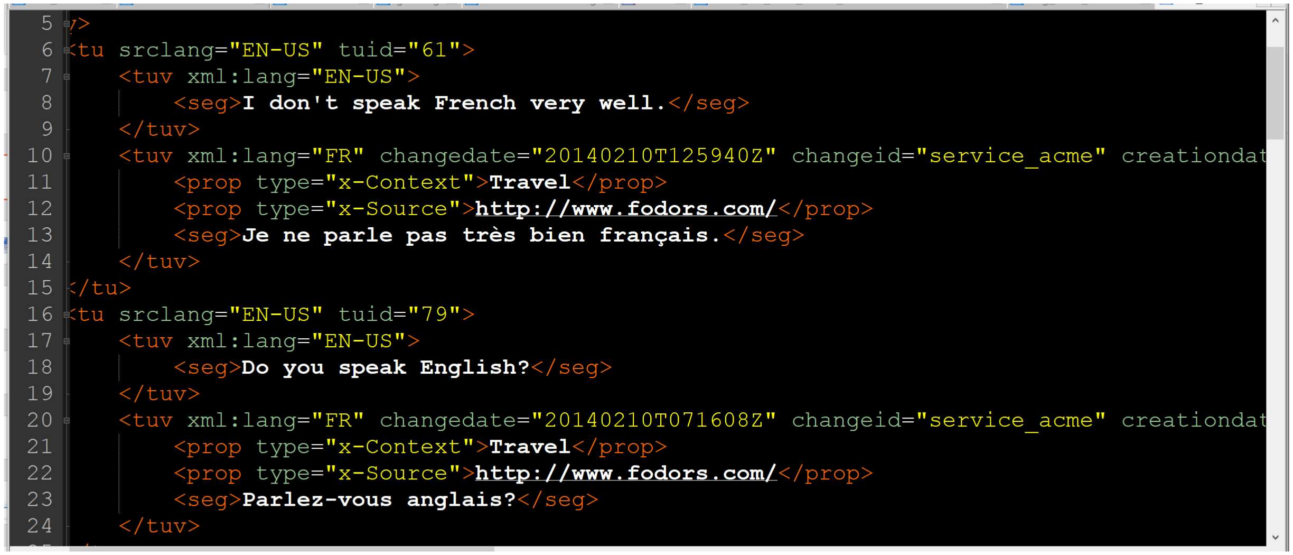
- Translators use saved translations so they can reduce future translation work (reference, fuzzy matching)
- Localization Engineers use TMs when they develop Enterprise TMS software
- Machine Translation Engineers use TMs when they train machine translation models

Summary:

Translation memories are a valuable asset for translators and organizations. Effective management of TMs is critical to leveraging their potential and reducing cost with expanding an organization’s global footprint.

What is a TMX

- TMX stands for Translation Memory eXchange, an XML file format



```
/>
<tu srclang="EN-US" tuid="61">
    <tuv xml:lang="EN-US">
        <seg>I don't speak French very well.</seg>
    </tuv>
    <tuv xml:lang="FR" changedate="20140210T125940Z" changeid="service_acme" creationdat
        <prop type="x-Context">Travel</prop>
        <prop type="x-Source">http://www.fodors.com/</prop>
        <seg>Je ne parle pas très bien français.</seg>
    </tuv>
</tu>
<tu srclang="EN-US" tuid="79">
    <tuv xml:lang="EN-US">
        <seg>Do you speak English?</seg>
    </tuv>
    <tuv xml:lang="FR" changedate="20140210T071608Z" changeid="service_acme" creationdat
        <prop type="x-Context">Travel</prop>
        <prop type="x-Source">http://www.fodors.com/</prop>
        <seg>Parlez-vous anglais?</seg>
    </tuv>
```

- For more information on the TMX file format:
  https://en.wikipedia.org/wiki/Translation_Memory_eXchange
  https://www.gala-global.org/tmx-14b

Types of Translation Memory Management

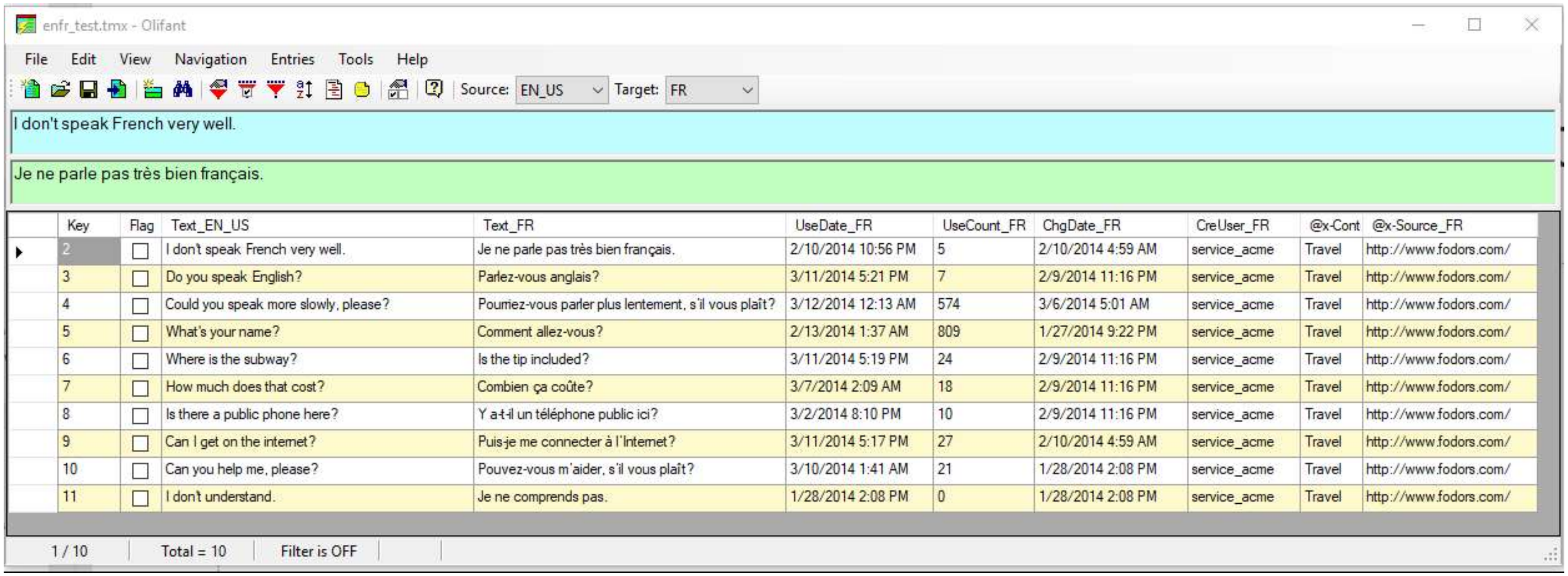
There are two basic types of TM Management:

Translation Management System (TMS)	Enterprise end-to-end solution. Very few free open source offerings. Often include desktop CAT tools.
Computer Assisted Translation Tools (CAT)	Various types of tools that deal with a specific stages of the translation process. Usually a desktop tool that will allow translators without direct access to a TMS to do work offline and submit their translations.

Open Source Desktop TM Management Tools

Okapi Framework - Olifant:

- http://okapi.sourceforge.net/Release/Olifant/Help/
- https://bitbucket.org/okapiframework/olifant/overview
- https://github.com/OkapiFramework/okapi-olifant



Source: EN_US Target: FR

I don't speak French very well.

Je ne parle pas très bien français.

	Key	Flag	Text_EN_US	Text_FR	UseDate_FR	UseCount_FR	ChgDate_FR	CreUser_FR	@x-Cont	@x-Source_FR
▶	2	□	I don't speak French very well.	Je ne parle pas très bien français.	2/10/2014 10:56 PM	5	2/10/2014 4:59 AM	service_acme	Travel	http://www.fodors.com/
	3	□	Do you speak English?	Parlez-vous anglais?	3/11/2014 5:21 PM	7	2/9/2014 11:16 PM	service_acme	Travel	http://www.fodors.com/
	4	□	Could you speak more slowly, please?	Pourriez-vous parler plus lentement, s'il vous plaît?	3/12/2014 12:13 AM	574	3/6/2014 5:01 AM	service_acme	Travel	http://www.fodors.com/
	5	□	What's your name?	Comment allez-vous?	2/13/2014 1:37 AM	809	1/27/2014 9:22 PM	service_acme	Travel	http://www.fodors.com/
	6	□	Where is the subway?	Is the tip included?	3/11/2014 5:19 PM	24	2/9/2014 11:16 PM	service_acme	Travel	http://www.fodors.com/
	7	□	How much does that cost?	Combien ça coûte?	3/7/2014 2:09 AM	18	2/9/2014 11:16 PM	service_acme	Travel	http://www.fodors.com/
	8	□	Is there a public phone here?	Y a-t-il un téléphone public ici?	3/2/2014 8:10 PM	10	2/9/2014 11:16 PM	service_acme	Travel	http://www.fodors.com/
	9	□	Can I get on the internet?	Puis-je me connecter à l'Internet?	3/11/2014 5:17 PM	27	2/10/2014 4:59 AM	service_acme	Travel	http://www.fodors.com/
	10	□	Can you help me, please?	Pouvez-vous m'aider, s'il vous plaît?	3/10/2014 1:41 AM	21	1/28/2014 2:08 PM	service_acme	Travel	http://www.fodors.com/
	11	□	I don't understand.	Je ne comprends pas.	1/28/2014 2:08 PM	0	1/28/2014 2:08 PM	service_acme	Travel	http://www.fodors.com/

1 / 10 Total = 10 Filter is OFF

Olifant Limitations:

- Hasn’t been maintained in four years
- Custom properties (meta-data) are not strongly typed which causes problems with sorting and filtering
- Doesn’t allow you to manage a repository of TMs. Just one single TMX file at a time.
- Limited import/export features
- Limited features for advanced users like localization engineers.

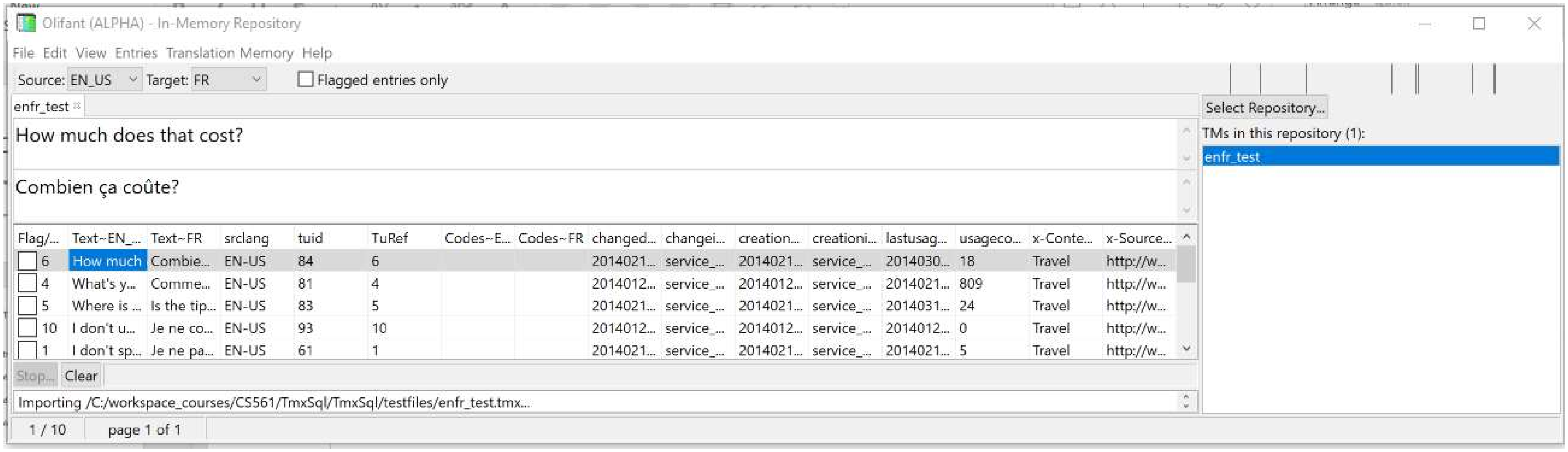
Other open source or free solutions:

- Google Translator’s Toolkit
  https://translate.google.com/toolkit

- Heartsome
  https://github.com/heartsome/tmxeditor8

Latest Olifant Development

- Allows management of multiple TMX files in a centralized repository
- In ‘ALPHA’ phase
- Last update was four years ago
- https://bitbucket.org/okapiframework/olifant



Olifant (ALPHA) - In-Memory Repository
File Edit View Entries Translation Memory Help
Source: EN_US Target: FR
□ Flagged entries only
enfr_test
Select Repository...
TMs in this repository (1):
enfr_test
Stop...
Clear

How much does that cost?

Combien ça coûte?

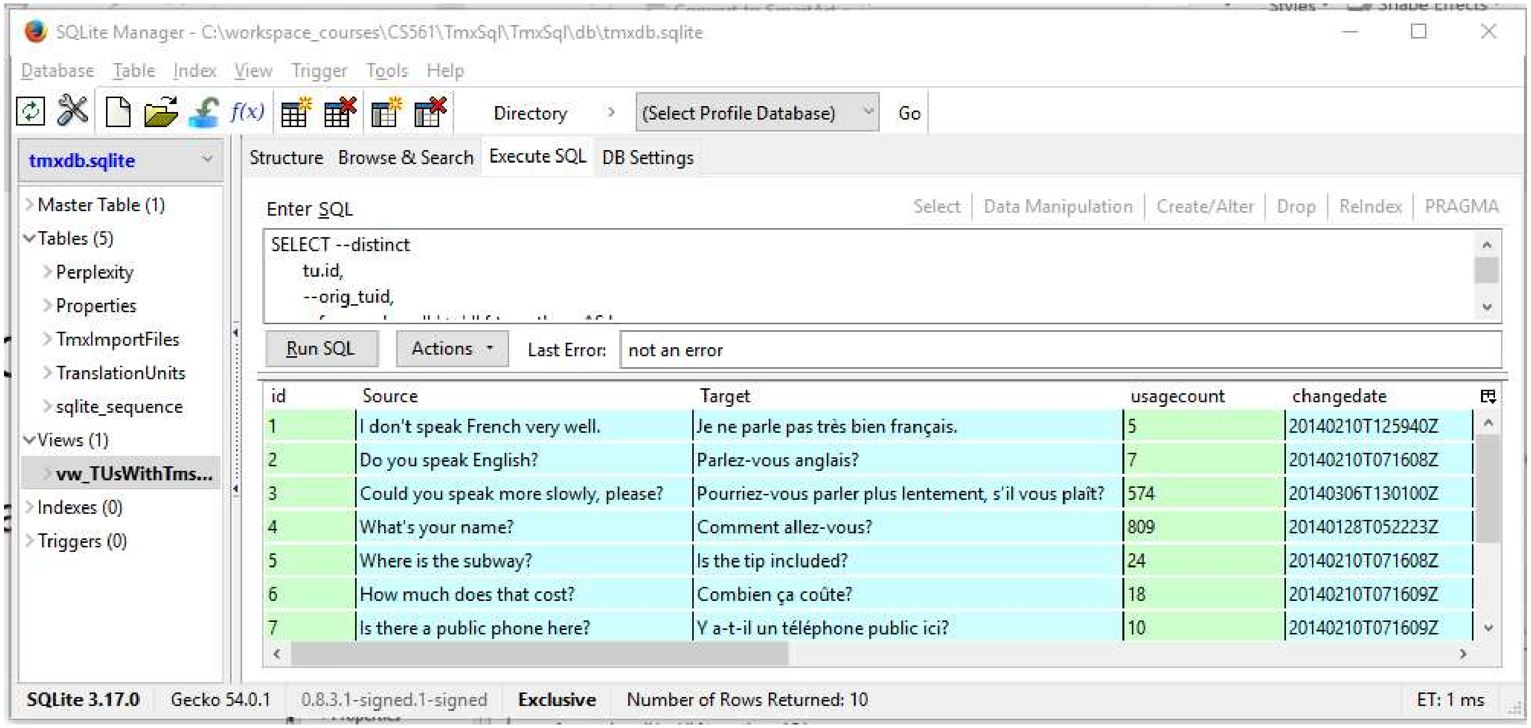
Flag/...	Text~EN_...	Text~FR	srclang	tuid	TuRef	Codes~E...	Codes~FR	changed...	changei...	creation...	creationi...	lastusag...	usageco...	x-Conte...	x-Source...
□ 6	How much	Combie...	EN-US	84	6			2014021...	service_...	2014021...	service_...	2014030...	18	Travel	http://w...
□ 4	What's y...	Comme...	EN-US	81	4			2014012...	service_...	2014012...	service_...	2014021...	809	Travel	http://w...
□ 5	Where is ...	Is the tip...	EN-US	83	5			2014021...	service_...	2014021...	service_...	2014031...	24	Travel	http://w...
□ 10	I don't u...	Je ne co...	EN-US	93	10			2014012...	service_...	2014012...	service_...	2014012...	0	Travel	http://w...
□ 1	I don't sp...	Je ne pa...	EN-US	61	1			2014021...	service_...	2014021...	service_...	2014021...	5	Travel	http://w...

Importing /C:/workspace_courses/CS561/TmxSql/TmxSql/testfiles/enfr_test.tmx...

1 / 10 page 1 of 1

Introducing: TmxSql

TmxSql uses a python script to import TMX files into a single Sqlite database where engineers can easily manage TMs using SQL.



```
SELECT --distinct
    tu.id,
    --orig_tuid,
```

Last Error: not an error

id	Source	Target	usagecount	changedate
1	I don't speak French very well.	Je ne parle pas très bien français.	5	20140210T125940Z
2	Do you speak English?	Parlez-vous anglais?	7	20140210T071608Z
3	Could you speak more slowly, please?	Pourriez-vous parler plus lentement, s'il vous plaît?	574	20140306T130100Z
4	What's your name?	Comment allez-vous?	809	20140128T052223Z
5	Where is the subway?	Is the tip included?	24	20140210T071608Z
6	How much does that cost?	Combien ça coûte?	18	20140210T071609Z
7	Is there a public phone here?	Y a-t-il un téléphone public ici?	10	20140210T071609Z

SQLite 3.17.0 | Gecko 54.0.1 | 0.8.3.1-signed.1-signed | Exclusive | Number of Rows Returned: 10 | ET: 1 ms

TmxSql Features:

- You can centralize your TMs in a single repository
- Allows localization engineers to use SQL to manage TMs
- Imports custom properties from TMX files
- Other than installing a Sqlite and a python script, no other software is required

TmxSql Limitations:

- Only works with bilingual corpora, not multilingual corpora
- Not suitable for translators or users that are not localization engineers
- Issues dealing with ambiguities related to custom properties in TMX

Self Assessment / Lessons Learned:

1. Defining the project goal was/is difficult
2. Spent a lot of time trying to join the Okapi Framework team.
3. Tried to get a too many birds with one stone:
   - A potential client wanted to pay me to work on something similar
   - I tried to make that work open source
   - Tried to make it part of the Okapi Framework
   - And… to use the Okapi Framework project as my class project
4. I spent less time adding new features to my project and more time on point #3.
5. Failed to drum interest by potential contributors and users.

Interested in Contributing?

TmxSql:

Steve Braich

stevebpdx@gmail.com

https://github.com/stevebpdx/TmxSql

Okapi Framework:

https://bitbucket.org/okapiframework