Построение технологического процесса с использованием ТМ для перевода сложных типов файлов

Демид Тишин,
бюро переводов «Окей»

Самара, 30-31 мая 2009 г.

Понятия презентации

- **Документ / файл оригинала** = текст + нетекстовая информация (расположение и форматирование текста, изображения и т.д. / теги и обозначения переменных)
- **Редактируемый документ / файл:**
  содержит текстовую информацию (в цифровом виде), отображаемую как символы, доступные для перезаписи
  примеры: doc, txt, dwg, dxf, cdr, pdf (некоторые),
- **Читаемый документ / файл:**
  содержит текстовую информацию (в цифровом виде), отображаемую как символы, доступные для чтения (копирования)
  примеры: pdf, djvu, dxf (некоторые)
- **Нередактируемый документ / файл:**
  Содержит текстовую информацию в виде изображения
  пример: jpg, bmp, pdf на основе изображения и т.д.
- **Текст** = исключительно текстовая информация, подлежащая переводу

Перевод без использования ТМ

- **Нередактируемый документ / файл оригинала:**
  1) перевод «с чистого листа»
  на выходе – 1 файл (перевод)
  Преимущество: быстрота
  Недостаток: нулевая обучаемость системы

  2) создание редактируемого файла
  (OCR, конвертеры и т.д.)
- **Редактируемый документ/файл оригинала:**
  перевод поверх текста оригинала
  на выходе – 2 разрозненных файла (оригинал и перевод)

Перевод с использованием TM

- Перевод всегда осуществляется внутри параллельного двуязычного текста → → оригинал всегда требуется в редактируемом виде → временные потери
- на выходе – min. 4 файла:
    1. оригинал
    2. перевод
    3. законченный параллельный текст
    4. обновленная база данных TM

дополнительно – итоговый глоссарий

дополнительно – терминологическая база

Техпроцесс для читаемых файлов

читаемый документ/файл

1. снятие защиты
2. копирование текста в программе-читателе / использование конвертера

редактируемый файл с упрощенным форматированием

3. подгрузка в проект TM-программы

двуязычный рабочий файл (ttx, djvpr и т.д.)

4. подключение TM, терминологической базы
5. анализ текста оригинала на совпадения с TM
6. расчет трудозатрат и цены
7. предперевод с подстановкой совпадений

частично переведенный двуязычный рабочий файл

Техпроцесс для читаемых файлов

(продолжение)

частично переведенный двуязычный рабочий файл

8. работа над текстом в интерфейсе TM-программы (перевод, редактирование, корректура и т.д.)
NB! Обновление базы TM
(в «живом» или в дискретном режиме)

8а. экспорт в простой формат файла

частично переведенный текст в виде двухколоночной таблицы rtf/html

8б. работа над текстом в интерфейсе текстового редактора (перевод, редактирование, корректура и т.д.)

Полностью переведенный текст в виде двухколоночной таблицы

8в. импорт обратно в TM-программу
NB! Обновление базы TM в дискретном режиме

Полностью переведенный двуязычный рабочий файл

Техпроцесс для читаемых файлов

(продолжение)

Полностью переведенный двуязычный рабочий файл

9. автоматизированный контроль качества
10. исправление ошибок
11. экспорт готового текста перевода / cleanup

редактируемый файл перевода с упрощенным форматированием

12. верстка в графическом редакторе или использование конвертера

читаемый или редактируемый файл перевода с полной нетекстовой информацией

Техпроцесс для читаемых файлов

- Поскольку текст в файле оригинала не может быть перезаписан, используются промежуточные редактируемые файлы с упрощенным форматированием (например, txt).
  Такие форматы файлов поддерживаются даже бесплатными ТМ-программами (напр., **OmegaT**)
- Экспорт рабочего файла ТМ-программы в простой формат – возможность только некоторых ТМ-программ
- При использовании экспорта в простой формат обновление ТМ происходит дискретно

Техпроцесс для редактируемых файлов

Редактируемый документ/файл

1. подгрузка в проект TM-программы

двуязычный рабочий файл (ttx, djvpr и т.д.)

2. подключение TM, терминологической базы
3. анализ текста оригинала на совпадения с TM
4. расчет трудозатрат и цены
5. предперевод с подстановкой совпадений

частично переведенный двуязычный рабочий файл

Техпроцесс для редактируемых файлов

(продолжение)

частично переведенный двуязычный рабочий файл

6. работа над текстом в интерфейсе TM-программы (перевод, редактирование, корректура и т.д.)
NB! Обновление базы TM
(в «живом» или в дискретном режиме)

6а. экспорт в простой формат файла

частично переведенный текст в виде двухколоночной таблицы rtf/html

6б. работа над текстом в интерфейсе текстового редактора (перевод, редактирование, корректура и т.д.)

Полностью переведенный текст в виде двухколоночной таблицы

6в. импорт обратно в TM-программу
NB! Обновление базы TM в дискретном режиме

Полностью переведенный двуязычный рабочий файл

Техпроцесс для редактируемых файлов
(продолжение)

Полностью переведенный двуязычный рабочий файл

7. автоматизированный контроль качества
8. исправление ошибок
9. экспорт готового текста перевода / cleanup

редактируемый файл перевода

10. проверка форматирования

редактируемый файл перевода

Техпроцесс для редактируемых файлов

- В данном случае, чем больше форматов поддерживает ТМ-программа, тем лучше.
- Поскольку при экспорте в редактируемый файл перевода из двуязычного рабочего файла возможно нарушение форматирования, для исправления ошибок желательно иметь программы-редакторы соответствующих форматов → удорожание себестоимости перевода
- Trados и Déjà Vu X не поддерживают форматы AutoCAD без дополнительной настройки

Типичные специфические сложности

- **Чертежи AutoCAD:**

  большое количество маленьких фрагментов текста (отдельные слова и обозначения) при экспорте (напр., программой dxf2txt) сохраняются в виде списка, т.е. одномерно, в то время как чертеж двумерный →
  → необходимость постоянно обращаться к исходному файлу при переводе и редактировании (исполнителю требуется программа-читатель!)

Типичные специфические сложности

- **Чертежи AutoCAD: извлечение текста**

```
№ док.
{8E6F}
Подп.
{8E6E}
Кол уч.
{8E6D}
Лист
{8E6C}
Изм.
{8E62}
СПЕЦИФИКАЦИЯ ТЕХНОЛОГИЧЕСКОГО ОБОРУДОВАНИЯ (НАЧАЛО)
{8E5F}
СПЕЦИФИКАЦИЯ ТЕХНОЛОГИЧЕСКОГО ОБОРУДОВАНИЯ (НАЧАЛО)
{8E5E}
г. САМАРА, МОСКОВСКОЕ ш. 23й км., с/х "КРАСНЫЙ ПАХАРЬ", ТРЦ "МЕГА"
{8E5D}
РЕСТОРАН "IL ПАТИО"
{8E51}
Изм.
{8E50}
Лист
{8E4F}
Кол уч.
{8E4E}
Подп.
{8E4D}
№ док.
{8E4C}
Дата
{8E4A}
Лист
{8E48}
5
{8E47}
ПОЯСНИТЕЛЬНАЯ ЗАПИСКА (ОКОНЧАНИЕ)
{8E45}
```

Типичные специфические сложности

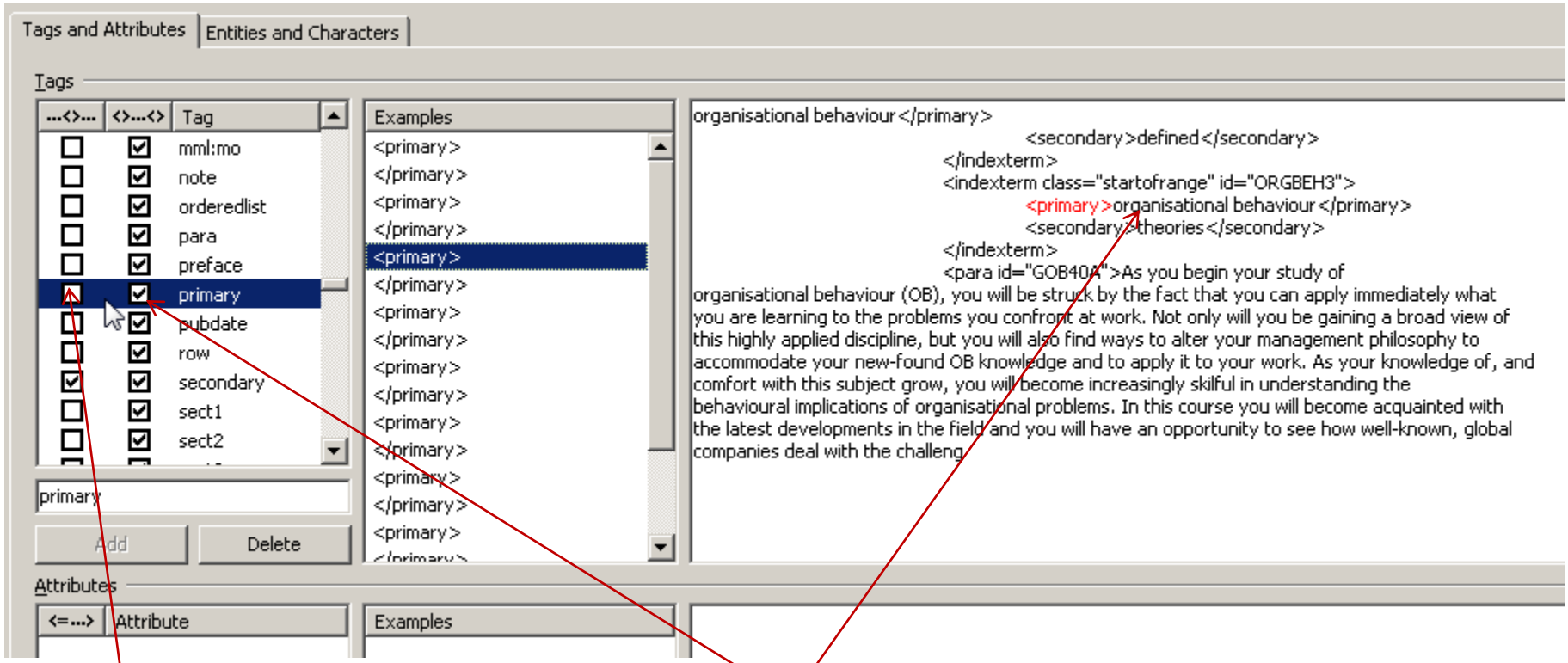
- **XML-файлы (*e**X**tensible **M**arkup **L**anguage*)***

  Нередко у заказчика отсутствует определитель фильтра (filter definition file), который требуется TM-программе для отделения текста для перевода от остального содержимого → определитель фильтра приходится создавать вручную

  ** текстовый формат, предназначенный для хранения структурированных данных*

Типичные специфические сложности

- **XML-файлы: создание фильтра**



Тег со всем содержимым не будет включаться в другие сегменты

Текст между открывающим и закрывающим тегом требует перевода

Типичные специфические сложности

- **Нередактируемые документы:**

  Если сроки жесткие, а документов мало, на создание редактируемого файла из нередактируемого (OCR или набор текста) не хватает времени → перевод и редактирование осуществляются в графическом редакторе → TM-программы использовать невозможно

  NB! Редактирование множества текстовых фрагментов в одном файле облегчается за счет использования векторных редакторов
  («Найти и заменить» - «Заменить все»)

  NB! Если документы однотипные и в большом количестве, а TM-программа обладает функцией автоподстановки перевода в целой группе файлов (напр., Déjà Vu X), перевод документов в редактируемый формат может быть оправдана

Спасибо за внимание!

Вопросы?