boost
your content
to success

©Festo SE & Co. KG, alle Rechte vorbehalten

Sehr frei nach Marie Kondo: Magic Cleaning – Aufräumen mit [itl]-clean

Wie die Firma Festo dem Datenschatz in ihren Translation Memories ein großes Stück nähergekommen ist

Success-Story

©Festo SE & Co. KG, alle Rechte vorbehalten

Wo ist der Schatz begraben?

Translation-Memory-Inhalte sind unbestreitbar ein Vermögenswert in international agierenden Unternehmen. Sie umfassen große Teile aller im Unternehmen vorhandenen Contents in verschiedenen Sprachen unter Berücksichtigung der Corporate-Wording-Vorgaben. Sie sind der Garant für effiziente und kostengünstige Übersetzungsprozesse. Sie sind ein heiliger Gral im Unternehmen. Sie sind „Big Data".

Aber ganz häufig sind TM-Daten unstrukturiert, veraltet und aufgebläht, nach dem Motto „Im Zweifel behalten", und ähneln mehr einem vollgestellten Dachboden als einer Schatzkiste.

Vollgestellte Dachböden haben mit großen TM-Beständen gemeinsam, dass man weiß, man sollte aufräumen, aber jeder Anlauf ist so abschreckend, dass man lieber wieder die Türe schließt.

Aber irgendwann kommt der Zeitpunkt, an dem man den Tatsachen nicht mehr ausweichen kann.

Man zieht um, man braucht den Platz für anderes oder will etwas machen aus den gelagerten Beständen. Sprich, man wechselt das System, der Speicherplatzbedarf wird zu groß und unhandlich oder man möchte den Datenbestand beispielsweise für das Training einer maschinellen Übersetzungsmaschine nutzen.

Dann muss man ran an die gesammelten Werke.

Kommt man allein nicht weiter, gibt es im Dachbodenbeispiel Umzugshelfer und Ordnungsberater, für die Bewältigung eines Content-Cleanups gibt es itl.

Mit dem Einzug von zugänglichen, bezahlbaren und vor allem effizienten Lösungen im Bereich der künstlichen Intelligenz und des Machine Learning in die Welt von Content und Sprachen haben wir bei itl nicht bei dem offensichtlichen Schwerpunkt Machine Translation aufgehört, sondern die zahlreichen Möglichkeiten in dem neuen Kompetenzbereich *intelligent solutions* zum Kernthema gemacht.

Dabei sind den Spezialisten Translation-Memory-Inhalte als Big Data schnell ins Auge gesprungen. Im engen Austausch mit den Kolleginnen und Kollegen aus dem Kompetenzbereich *global translation* sowie mit unseren langjährigen Kunden sind die „Pain Points" aus der Lokalisierung den Software-Entwicklern schnell klar geworden.

©Festo SE & Co. KG, alle Rechte vorbehalten

Festo und die Schatztruhe auf dem Dachboden

Ein Unternehmen, das mit einem professionellen internen Lokalisierungsteam schon lange selbst am Puls der Zeit und an Prozessoptimierungen arbeitet, ist die Firma Festo SE & Co. KG. Wir schätzen uns glücklich, dass wir Festo bereits seit mehr als zehn Jahren zu unseren Kunden zählen dürfen. Festo fordert uns und spornt uns an, innovative Lösungen vorzuschlagen, einzubringen, auszuprobieren und diese dann in einer partnerschaftlichen Zusammenarbeit in die Realität umsetzen.

Das jüngste Beispiel dieser Zusammenarbeit dürfen wir eine Erfolgsgeschichte nennen und der Protagonist ist [itl]-clean.

Die Firma Festo stand, wie viele andere international erfolgreiche Unternehmen, vor der Herausforderung, einen prall gefüllten Datenbank-Speicher auszumisten, mit der Angst, wertvolle Inhalte womöglich wegzuschmeißen und dem Gefühl, vor einer Sisyphus-Aufgabe zu stehen, die Zeit, Geld und Nerven kostet.

Die Daten im Translation Memory waren teils über 20 Jahre alt und hatten bereits einen CAT-Tool-Wechsel hinter sich. Damals waren bereits Merge-Dateien mit unterschiedlichen Ständen und von unterschiedlichen Dienstleistern mit migriert worden, weil trotz bester Absichten die Möglichkeiten einer effizienten, aber dennoch sinnvollen Datenbereinigung in der Vergangenheit ziemlich eingeschränkt waren.

Daneben entwickelte sich die Terminologie zu einer immer zentraleren Komponente im Lokalisierungsprozess, mit großen Investitionen in deren Aktualisierung. Dies wiederum erhöhte die Diskrepanz zu einer veralteten Translation-Memory-Datenbank.

Im Herbst vergangenen Jahres wurde im regelmäßigen Austausch mit dem Festo-Lokalisierungsteam bei itl unter Federführung der Teamleiterin Stephanie Morjan auf die Fortschritte in der Entwicklung eines TM-Bereinigungstools hingewiesen, das bei Festo sinnvoll eingesetzt werden könnte, um die auf beiden Seiten bekannten Probleme in den Griff zu bekommen. Parallel dazu fanden verschiedene Informationsveranstaltungen zum Thema durch Bastian Heilemann, Leiter Software Development (aka *intelligent solutions*), bei itl statt und bald die ersten Termine zur Abstimmung der Erwartungshaltung mit Festo.

Die Anforderungen von Festo waren klar und anspruchsvoll:

- Die Zeitschiene für die Bereinigung sollte sich in einem akzeptablen Rahmen bewegen, um rasch mit den Daten weiterarbeiten zu können, aber v.a. auch, um kein großes Delta an neuem Content während der Bereinigungsphase auflaufen zu lassen.
- Es sollte möglich sein, die aussortierten Inhalte bei Bedarf sichten zu können, um die latente Angst zu beruhigen, man könnte Inhalte aussortieren und wegwerfen, die man noch mal brauchen würde.
- Die Bereinigung der Bestandsdaten musste kostentechnisch vertretbar und ressourcenschonend sein. Eine manuelle Bearbeitung kam nicht in Frage.
- Das mit der Bereinigung erreichte Qualitätsniveau der Inhalte sollte in einem kontinuierlichen Qualitätssicherungsprozess gehalten werden können.

Mit [itl]-clean das richtige Werkzeug in Händen

Das Tool zur automatisierten und nachhaltigen Bereinigung großer mehrsprachiger Content-Bestände, das bei itl seit einiger Zeit im Entstehen war, ist [itl]-clean, ausgelegt als eine skalierbare Cloud-Lösung.

Mit deren Hilfe werden kurz gesagt Textsegmente aus dem Datenbestand entfernt, die automatisiert erkennbare Fehler enthalten, und diese werden dann in eine separate Datenbank ausgespielt.

Zusammen mit den versierten Ansprechpartnern der Translation and Content Services bei Festo haben wir [itl]-clean den Feinschliff verliehen und das Tool schließlich zur großen Zufriedenheit aller Beteiligten in großem Umfang eingesetzt.

Die Vorteile einer solchen Bereinigung zeigen sich einerseits im Datenmanagement reduzierter Datenmengen, die damit weniger Speicherplatz benötigen und weniger Migrationsaufwände verursachen.

Und andererseits in der inhaltlichen Optimierung der Datenqualität: Saubere und eindeutige Datenbestände verbessern die Übersetzungsqualität, vereinfachen die Qualitätssicherung und reduzieren Fehlerquellen und damit negative Rückmeldungen und aufwendiges Beschwerdemanagement.

Was genau kann [itl]-clean und was wurde in dem Projekt mit Festo umgesetzt und hat die Auftraggeber zu wahren Lobeshymnen verleitet?

Ein paar technische Daten müssen sein

[itl]-clean basiert auf einem Set aus skalierbaren KI-Algorithmen, die auf objektiv erkennbare und messbare Eigenschaften von mehrsprachigem Content abzielen. Die Auswahl, Konfiguration und Kombination derselben ist Ergebnis der klugen Kombination des langjährigen itl-Fachwissens in der Lokalisierungsindustrie und Best Practices namhafter F&E-Institutionen im Lokalisierungsbereich mit dem aktuellen Know-how im Bereich Software-Entwicklung im itl-Kompetenzbereich *intelligent solutions*.

Nachstehend eine Übersicht über die aktuell eingesetzten Algorithmenbereiche mit ein paar typischen Beispielen:

Sprache	Terminologie	Übersetzung Heuristik	Formale Fehler	Metadaten
Spracherkennung: Translation Unit enthält nicht die erwartete Quell- oder Zielsprache	Mehrsprachige Terminologieprüfung gemäß Termbank oder Glossar	Segmentlängen-unterschiede zwischen Quelle und Ziel	Fehlende erwartbare Zeichen (Satzzeichen, fehlende Klammer)	Veraltete Segmente (Eintragsdatum)
	Granulare Konfiguration anhand Identifikation, z. B. Identifikation von verbotenen Termen, bereichsspezifischen Termen, Termen ohne Zielsprache …	Einwort-Segmente oder überlange Segmente ungeeignet für MT-Engine-Training	Leerzeichen-Normalisierung	Einträge bestimmter Bereiche, Autoren etc.
		Leere Segmente		
		Segmente nur mit Symbolen oder Zeichen		
		Tag-lastige Segmente		

Die Vorgehensweise ist folgendermaßen:

- Algorithmen werden definiert
  - > Alle Prüfungen sind individuell anpassbar je nach Zielsetzung (Migration oder MT-Training) und Sprache (Satzzeichen sind nicht immer identisch in Quelle und Ziel) usw.
- Prüfungen werden über alle Algorithmen auf einmal oder nacheinander durchgeführt
- Translation Units, die nicht entsprechen, werden entfernt bzw. in eine neue Datenbank ausgespielt
- Das Ergebnis ist eine bereinigte Datenbank und die Datenbank(en) mit dem Ausschuss

Die Prüfungen funktionieren derzeit bereits für einen Großteil der europäischen und asiatischen Sprachen, einschließlich einiger Ländervarianten.

```
<tu creationdate="20190429T111943Z" changedate="20190429T111943Z" creationid="49898926230">
    <prop type="user-defined">quality:100</prop>
    <prop type="user-defined">storing mode:1</prop>
    <prop type="user-defined">uncontrolled:0</prop>
    <prop type="user-defined">context left:7091</prop>
    <prop type="user-defined">context right:7092</prop>
    <prop type="user-defined">basestyle:&lt;cf extstyle="no"&gt;</prop>
    <prop type="Att::Fachgebiete">General</prop>
    <prop type="Att::Relationen">DE Quellsprache</prop>
    <prop type="Att::Text-Bereich">Trilogien</prop>
    <prop type="Att::Projekte">30401_DNO_Lambrecht, Leonhard_LMR_90157601_901939310037_ES-ES</prop>
    <tuv xml:lang="en-GB"><seg>Wir haben den richtigen Ansatz für Sie.</seg></tuv>
    <tuv xml:lang="es-ES"><seg>Tenemos el enfoque correcto para usted</seg></tuv>
</tu>
```

Abb.: Beispiel Spracherkennung: Deutscher Text ist unter einem als UK-Englisch ausgezeichneten Eintrag abgespeichert

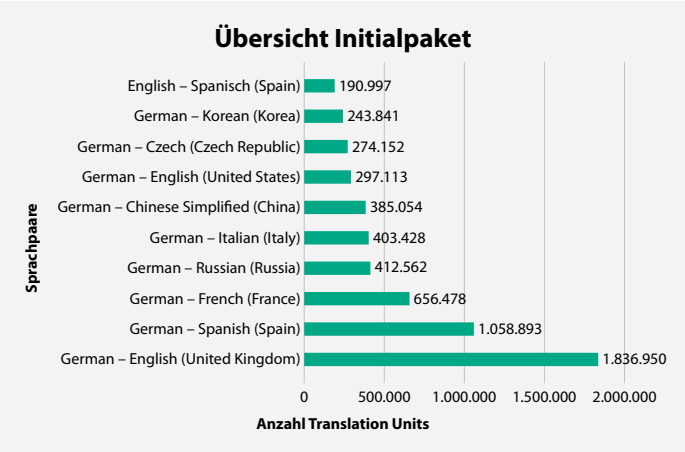
Das hat Festo überzeugt

Nach dem Austausch der Erwartungshaltungen und verschiedener Tests fiel bei Festo schließlich die Entscheidung für [itl]-clean. Neben der Voraussetzung, dass die Bereinigung wie vorgesehen funktionieren würde, waren die Hauptkriterien zugunsten des Tools von itl

- die separate Datenbank mit den aussortierten Daten
- der problemlose Re-Import der bereinigten Daten (das Tool verändert die Struktur der tmx-Datei in keiner Weise)
- der Report über das Gesamtergebnis

Das Projekt in Zahlen und Fakten

Im ersten abgeschlossenen Durchgang wurden schließlich 10 Sprachkombinationen mit insgesamt 5,8 Millionen Translation Units bereinigt.



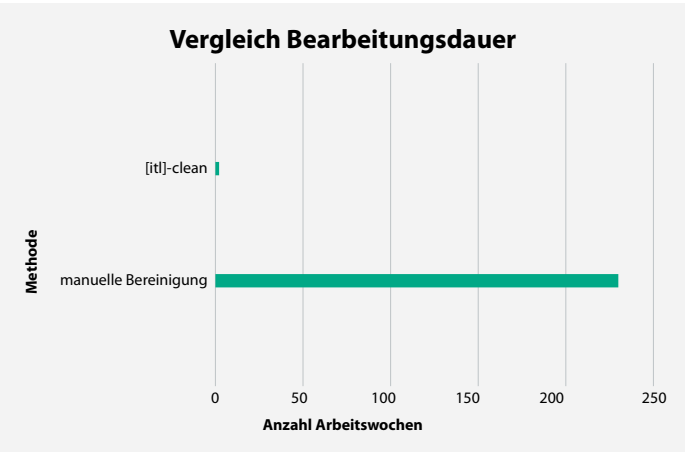
Bearbeitungszeit

Das oben genannte Volumen von insgesamt 5.759.448 Translation Units konnte innerhalb von zwei Wochen vollständig automatisiert geprüft werden. Manueller Aufwand von wenigen Stunden entstand nur im Vorfeld für die Datensichtung und Abstimmung über die zu prüfenden Parameter.

Das heißt, während der Bereinigungsphase war nur ein geringes Delta an neuem Content hinzugekommen, das mittels Datumangabe identifiziert und somit bei der nächsten Bereinigung miteingeschlossen werden kann.

Für eine manuelle Überprüfung unter der idealtypischen Annahme, dass jeweils ein Übersetzer pro Sprache parallel ausschließlich an dieser Aufgabe arbeiten würde, müsste man wohl grob vier Jahre rechnen.

Im Übrigen sind die Software-Architekten bei itl mit der aktuellen Performance von [itl]-clean noch lange nicht zufrieden und tüfteln bereits an schnelleren Abläufen.

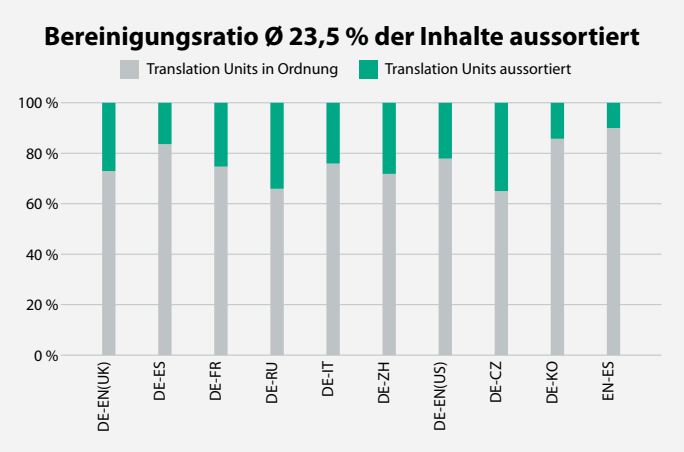


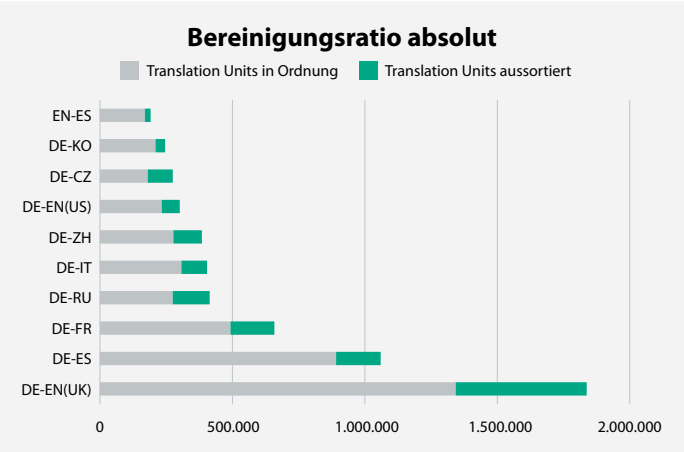
Ergebnisreport

Die spannendste Frage nach Abschluss der Bereinigung war natürlich: Wie viel wurde aussortiert und vor allem was?

Wie viel?

Angesichts der umfangreichen und teils auch alten Inhalte war der durchschnittliche „Ausschluss“ von 23,5 % des Contents vollkommen nachvollzieh- und erwartbar. Dabei lagen die Werte bei den einzelnen Sprachkombinationen zwischen 10 % beim kleinsten Translation Memory Englisch-Spanisch und 34 bzw. 35 % bei Deutsch-Russisch und Deutsch-Tschechisch.





Was?

Selbstverständlich legte Festo bei diesem ersten umfangreichen Einsatz von [itl]-clean Wert darauf, das Ergebnis auch inhaltlich zu überprüfen. Es wurden daher Stichproben aus allen Datenbanken genommen, in denen die automatisiert aussortierten Translation Units abgespeichert waren.

Das Ergebnis entsprach genau den Erwartungen. Wobei falsch zugeordnete Sprachen und Terminologiefehler die bei weitem am häufigsten vertretenen Fehlerkategorien waren.

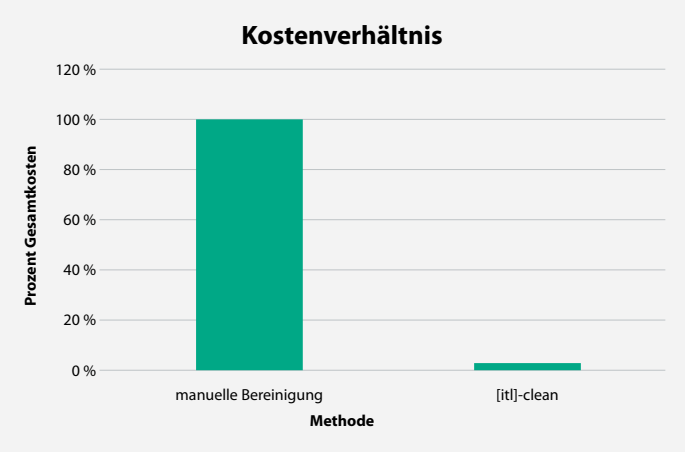
Die Schrecksekunde

Dann erreichte uns bei itl plötzlich die Rückmeldung, Festo-Lektoren hätten eine Anzahl falscher Translation Units im bereinigten Material gefunden, die falsche Terminologie enthielten. Das hat selbstverständlich die itl-Entwickler auf den Plan gerufen, die sich umgehend an die Ursachenforschung machten. Was, wenn ein Fehler im System den ganzen Erfolg in Frage stellte?

Nach wenigen Stunden konnte glücklicherweise Entwarnung gegeben werden: Es war der Mensch und die Anweisung, die er dem System gegeben hatte. Einwort-Segmente sollten standardmäßig nicht „bereinigt" werden, weil durchaus gewünscht sein kann, dass im Quelltext derselbe Wortlaut steht wie im Zielsegment; denken Sie nur an „Server". Als Translation Unit mit nur einem Wort wäre dieses Segment als „falsche" Sprache aussortiert worden. Dank der hohen Performanz von [itl]-clean sind Anpassungen der Parameter und Testläufe einfach möglich.

Kostenrahmen

Analog zum obigen Vergleich der Bearbeitungszeiten ist es nicht schwer, auch entsprechende Kostenvorteile abzuleiten. Selbst bei einem niedrig angesetzten Wortpreis für die Revision durch humane Übersetzer kommt man angesichts der Menge der TM-Inhalte sogleich auf utopische Summen, die in keinerlei Kosten-Nutzen-Verhältnis mehr stehen.



Nachhaltigkeit

Mit dem dargestellten Durchbruch in der Effizienzsteigerung bei der Bereinigung großer Datenbanken ist es nunmehr auch möglich, diese Prozesse in einen regelmäßigen Qualitätssicherungszyklus zu implementieren. Dadurch entstehen keine derart großen Bestandsdatenmengen mehr, die einer Grundreinigung bedürfen, die Deltas zwischen den Bereinigungszeiten werden noch kleiner und die Aufwände werden zusehends geringer.

Wie geht es weiter?

Festo hat inzwischen weitere Translation-Memory-Datenbanken zur Bereinigung geschickt, während die „sauberen" bereits für die laufenden Übersetzungsprojekte im Einsatz sind.

Mit dem jetzt aktuellen und leichteren Daten-Gepäck kann das Unternehmen die nächsten Schritte in die Nutzung des Datenschatzes Translation Memory gehen.

Bei itl sind die Experten aus *intelligent solutions* und *Localization Engineering* den nächsten Trends auf der Spur, die Kunden und itl helfen, uns die neuen Technologien aus der Welt von KI, Cloud Native und System-Konnektoren zunutze zu machen. Dann haben Übersetzungsbeauftragte, Projektmanager, Übersetzer und Co. mehr Zeit, sich der Sprache, den Prozessen und der Qualität zu widmen, anstatt Daten von Hand zu sortieren, zu verschieben und aufzubereiten.

©Festo SE & Co. KG, alle Rechte vorbehalten

Über Festo:

Die Festo Gruppe – ein Global Player in der Digitalisierung. Im Zuge der Digitalisierung richtet sich Festo als weltweit führender Anbieter von Automatisierungstechnik und technischer Bildung mit seinen Produkten und Services auf die smarte Produktion der Zukunft aus. Dabei setzt das Unternehmen auch auf künstliche Intelligenz und maschinelles Lernen.

Seit über 60 Jahren ist das 1925 gegründete unabhängige Familienunternehmen mit Sitz in Esslingen a.N. Impulsgeber in der Automatisierung und hat sich mit einem einzigartigen Angebot zum Weltmarktführer in der technischen Aus- und Weiterbildung entwickelt. 300.000 Kunden in der Fabrik- und Prozessautomation weltweit vertrauen auf die pneumatischen und elektrischen Antriebslösungen des Unternehmens.

Über itl:

Seit 1982 helfen wir Herstellern weltweit dabei, mit ihren Technischen Dokumentationen gegenüber ihren Kunden zu brillieren. Als Full-Service-Anbieter haben wir dabei nicht nur das fertige Produkt in Text und Bild (gedruckt und digital) im Auge, sondern auch die dahinter liegende Infrastruktur, Software und Prozesse. Wir helfen Ihnen bei der Optimierung aller Aspekte – immer unter der Maßgabe, Ihnen höchste Qualität zu bieten.

Wir sind ein verlässlicher Partner für unsere Kunden und verbinden klassische Technische Dokumentation mit dem Start-up-Spirit der Development-Szene. Unsere Antworten auf die heutigen Anforderungen der Industrie sind:

- Bildorientierte und digitale Dokumentationskonzepte
- Maschinelle Übersetzung
- [i]-match, unsere selbstentwickelte Language Management Plattform
- Entwicklung automatisierter Schnittstellen zwischen Kunde und Dienstleister

Unsere Kernkompetenzen liegen in Übersetzung, Redaktion, Grafikerstellung, Publishing und Consulting, wobei insbesondere Wert auf die Optimierung und Automatisierung der Prozesse im Dokumentlebenszyklus gelegt wird.

itl Institut für technische Literatur AG
Elsenheimerstraße 65-67, D-80687 München
+49 89 892623-610
vertrieb@itl.eu

40 JAHRE [itl]